Challenges and Opportunities for Knowledge Organization in the Digital Age

Proceedings of the Fifteenth International ISKO Conference
9-11 July 2018
Porto, Portugal

Organized by
International Society for Knowledge Organization (ISKO),
ISKO Spain and Portugal Chapter
University of Porto – Faculty of Arts and Humanities
Research Centre in Communication, Information
and Digital Culture (CIC.digital) – Porto

Edited by
Fernanda Ribeiro
Maria Elisa Cerveira
Challenges and Opportunities for Knowledge Organization in the Digital Age
Editorial Support:
Raquel Graça

Predocumentation:

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

© Ergon – ein Verlag in der Nomos Verlagsgesellschaft, Baden-Baden 2018
This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in other ways and storage in databanks. Duplication of this publication or parts thereof is only permitted under the provisions of the German Copyright Law, a copyright fee must always be paid.
Cover Design: Jan von Hugo
www.ergon-verlag.de

ISBN 978-3-95650-420-4 (Print)
ISBN 978-3-95650-421-1 (ePDF)
ISSN 0938-5495
# Table of Contents

**Introduction**

*Fernanda Ribeiro*-------------------------------------------------------------------------------------------------------------------------15

**Keynote Address**

Supporting truth and promoting understanding: Knowledge Organization and the curation of the infosphere  
*David Bawden*-------------------------------------------------------------------------------------------------------------17

**Foundations and Methods for Knowledge Organization**

Among dictionaries, reference works and concept systems: can the terminology in International Standard ISO 5127:2017 contribute to Knowledge Organization?  
*Axel Ermert*---------------------------------------------------------------------------------------------------------------29

An Analysis of the theoretical and practical application of Diplomastics to archival description in Knowledge Organization  
*Natalia Bolfarini Tognoli, Ana Célia Rodrigues*--------------------------------------------------------------------------------43

Applied Knowledge Organization and the history of the world  
*Rick Szostak*-------------------------------------------------------------------------------------------------------------52

Approaches to the concepts of exhaustivity and specificity in ISKO International meeting proceedings: 2000-2017  
*Maria da Graça Simões, Daniel Martínez-Ávila, Blanca Rodríguez-Bravo, Patricia de Almeida, Isadora Victorino Evangelista*-------------------------------------------------------------------------------------------------58

Articulating a cultural research program for Knowledge Organization Systems  
*Gregory H. Leazer, Robert Montoya, Jonathan Furner*-----------------------------------------------------------------------66

Aspects regarding the notion of subject in the context of different theoretical trends: teaching approaches in Brazil  
*Brisa Pozzi de Sousa, Cristina Dotta Ortega*--------------------------------------------------------------------------------74

Automatic indexing and ontologies: the consistency of research chronology and authoring in the context of Information Science  
*Maria da Graça Simões, Luís Miguel Machado, Renato Rocha Souza, Maurício Barcellos Almeida, António Tavares Lopes*----------------------------------------------------------------86
<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big data, Knowledge Organization and decision making: opportunities</td>
<td>Amos David, Nadine Ndjock</td>
</tr>
<tr>
<td>and limits</td>
<td></td>
</tr>
<tr>
<td>Classification and Knowledge Organization Systems: ontologies and</td>
<td>Thiago Henrique Bragato Barros, Daniel Libonati Gomes</td>
</tr>
<tr>
<td>archival classification</td>
<td></td>
</tr>
<tr>
<td>Classification of photographs: methodological concepts in Archival</td>
<td>Ana Cristina de Albuquerque, Luciane de Fátima Beckman Cavalcante</td>
</tr>
<tr>
<td>Science, Library Science and Museum Science</td>
<td></td>
</tr>
<tr>
<td>Comparing the use of full text search between a conventional IR</td>
<td>Edson Marchetti da Silva</td>
</tr>
<tr>
<td>System and a DBMS</td>
<td></td>
</tr>
<tr>
<td>Computer-assisted checking of conceptual relationships in a large</td>
<td>Decio Wey Berti, Jr., Gercina Lima, Benildes Maculan, Dagobert Soergel</td>
</tr>
<tr>
<td>thesaurus</td>
<td></td>
</tr>
<tr>
<td>The Contribution of Semiotics to Knowledge Organization for music</td>
<td>Camila Monteiro de Barros, Lígia Maria Arruda Café, Audrey Laplante</td>
</tr>
<tr>
<td>information</td>
<td></td>
</tr>
<tr>
<td>Critical questions for big data approach in knowledge representation</td>
<td>Lala Hajibayova, Athena Salaba</td>
</tr>
<tr>
<td>and organization</td>
<td></td>
</tr>
<tr>
<td>Developing a field of knowledge through bibliography: art history in</td>
<td>Giulia Crippa</td>
</tr>
<tr>
<td>the 16th century</td>
<td></td>
</tr>
<tr>
<td>Different parameters for Knowledge Organization in archives</td>
<td>Sonia Troitiño</td>
</tr>
<tr>
<td>Document representation with images: an experimental milestone</td>
<td>Marcilio de Brito, Widad Mustafa El Hadi, Maja Žumer, Simone Bastos</td>
</tr>
<tr>
<td></td>
<td>Vieira, Marcos de Brito</td>
</tr>
<tr>
<td>Domain analysis of scientific production in Information and</td>
<td>Fábio Mosso Moreira, Ely Francina Tannuri de Oliveira, Ricardo César</td>
</tr>
<tr>
<td>Communication Technology in the context of small farmers</td>
<td>Gonçalves Sant’Ana</td>
</tr>
<tr>
<td>Epistemic communities, domain analysis, and Kuhn: dialogs and</td>
<td>Daniel Martinez-Ávila, Isadora Victorino Evangelista, Maria da Graça</td>
</tr>
<tr>
<td>intersections in Knowledge Organization</td>
<td>Simões, José Augusto Chaves Guimarães</td>
</tr>
<tr>
<td>Epistemic LOCI: linguistic and critical meta-methodology in Knowledge</td>
<td>Tatiana de Almeida, Gustavo Silva Saldanha</td>
</tr>
<tr>
<td>Organization</td>
<td></td>
</tr>
<tr>
<td>Epistemological challenges in Knowledge Organization in the digital</td>
<td>Rosa San Segundo, Maria Adelina Codina-Canet</td>
</tr>
<tr>
<td>age</td>
<td></td>
</tr>
</tbody>
</table>
Facet: itself a multifaceted concept
Michèle Hudon, Alexandre Fortier ........................................... 204

Feasibility of implementing PRESSoo model in organizing Persian serials
Negin Shokrzadeh Hashtroudi, Mohsen Haji Zeinolabedini ........................................... 212

Foundations and methods for Knowledge Organization in European iSchools courses
Olívia Pestana ........................................................................ 224

A Framework to represent variables and values in Social Science research data sets to support data curation and reuse
Guangyuan Sun, Christopher S. G. Khoo ........................................... 231

The Genesis of Documentation in Brazil: Manoel Cicero Peregrino da Silva and Paul Otlet (1900-1924)
Carlos Henrique Juvêncio, Georgete Medleg Rodrigues ........................................... 240

Global Knowledge Organization, “super-facets” and music: universal music classification in the digital age
Deborah Lee, Lyn Robinson, David Bawden ........................................... 248

Historical ambiguity: a lens for approaching outdated terms
Lilium Rajan........................................................................ 256

How Knowledge Organization helped to shape the emerging field of Terminology in Canada
Lynne Bowker ........................................................................ 265

Influence of metatheoretical research in Knowledge Organization
Paula Carina de Araújo, Joseph T. Tennis ........................................... 273

Information and argument structures in Sociology research abstracts
Wei-Ning Cheng, Christopher S. G. Khoo ........................................... 282

Information and knowledge ecology: a field for research in Knowledge Organization
Wieslaw Babik ........................................................................ 290

Information from jussive processes
Elliott Hauser ........................................................................ 300

Intellectual history, history of ideas, and subject ontogeny
Joseph T. Tennis ........................................................................ 308

Is the massive incorporation of e-books into university libraries devaluing the technical processes related to the assigning of subject headings and classification codes?
Isidoro Gil-Leiva, Mariângela Spotti Lopes Fujita, Pedro Manuel Díaz Ortuño, Daniela Majorie dos Reis........................................... 314
Knowledge Organization in editorial policies for titles, abstracts and keywords in JCR-indexed journals: an exploratory study in the areas of Information and Communication Sciences
Mariângela Spotti Lopes Fujita, María-del-Carmen Agustín-Lacruz, Ana Lúcia Terra

Knowledge Organization in the health field: an ontology project to improve the information retrieval process
Claudio Jose Silva Ribeiro, Diones Ramos da Silva

Literary warrant-based approach to organize KO terminology: criteria and method
Mario Bartiê, Mirtha Rauch, Ana Inês Brozia, Micaela Morales

Medical ontology: Siddha System of Medicine
Hemalata Iyer, K.S. Raghavan

The Modern and its impact on models of information organization in Brazil: the decline of the National Library and the rise of the National Book Institute (1930-1954)
Carlos Henrique Juvêncio, Georgete Medleg Rodrigues

The Onomasiological approach and the function of definitions in the elaboration of domain models in ontologies
Maria Luiza de Almeida Campos, Hagar Espanha Gomes

OntoM4IS: ontology reuse method for Information Science
Helder Noel Monteiro Firmino, Gercina Ângela de Lima

Parsimony in Biological and Colon Classifications
Robert D. Montoya

The Penumbra-line: Ranganathan’s journeys and the genesis of the APUPA pattern
Luca Giusti

Photography as a legitimate technique for domain analysis in Knowledge Organization
Eva Hourihan Jansen

Reconstructionism: a comparative method for viewpoint analysis and indexing using the example of Kohlberg’s moral stages
Michael Kleineberg

A Relation typology in Knowledge Organization Systems: case studies in the research data management domain
Jian Qin

Reorganization of knowledge in technical documents for the information system: the extraction of logico-cognitive information from heterogeneous administrative data
Omar Larouk, Mabrouka El Hachani

The Retrieval power added by subject indexing to bibliographic databases
Philip Hider
Risks and requirements of an applied Information Science research framework in healthcare
Fernanda Gonçalves, Gabriel David

Semantic warrant, cultural hospitality and knowledge representation in multicultural contexts: experiments with the use of the EuroVoc and UNBIS thesauri
Roger de Miranda Guedes, Maria Aparecida Moura

Solid Foundations and some secondary assumptions in the design of bibliographic metadata: toward a typology of complementary uses of metadata
Athena Salaba, Joseph T. Tennis

The Spirit of inquiry’s power to influence in 21st-century KO research: Jesse Shera and Margaret Egan
José Augusto Chaves Guimarães, Maria Claudia Cabrini Gracio, Daniel Martinez-Ávila, Rodrigo de Sales

Truth, relevance, and justice: towards a veritistic turn for KO
Jonathan Furner

Verbal protocols in Brazilian Information Science: a perspective from indexing studies
Paula Regina Dal’ Evedove, Roberta Cristina Dal’ Evedove Tartarotti, Mariângela Spotti

Who is Tesauro? The man, words and things
Gustavo Silva Saldanha, Naira Christofoletti Silveira, Giulia Crippa and Tatiana de Almeida

Wikipedia categories in research: towards a qualitative review of uses and applications
Jesús Tramullas, Ana I. Sánchez-Casabón, Piedad Garrido-Picazo

Interoperability towards Information Access

A Comparative analysis and evaluation of bibliographic ontologies
Maria Teresa Biagetti

A Conceptual model for an OWL ontology to represent the knowledge of transmedia storytelling
Juan Antonio Pastor Sánchez, Tomás Saorín Pérez

Connecting KOSs and the LOD Cloud
Rick Szostak, Andrea Scharnhorst, Wouter Beek, Richard P. Smiraglia

Converting UDC to BCC: comparative approaches to interdisciplinarity
Richard P. Smiraglia, Rick Szostak
Culturally relevant relationships: publishing and connecting digital objects in collections of archives, libraries, and museums over the Web
Carlos H. Marcondes ---------------------------------- 539

Derivative interpretation of biographical sketches (bios) supporting innovative information
Marcia Lei Zeng, Sophy Shu-Jiun Chen ---------------------------------- 549

Extending the scope of library discovery systems via hashtags
Louise F. Spiteri ---------------------------------- 557

An FRBR-based approach for transforming MARC records into linked data
Ya-Ning Chen ---------------------------------- 565

From uniform identifiers to graphs, from individuals to communities: what we talk about when we talk about linked person data
M. Cristina Pattuelli ---------------------------------- 571

Identifying semantic characteristics of user interaction datasets through application of a data analysis
Fernando de Assis Rodrigues, Pedro Henrique Santos Bisi, Ricardo César Gonçalves Sant’Ana ---------------------------------- 581

Image organization on the Web: an analysis from the perspective of cultural heritage of rural farms in Brazil
Luciana de Souza Gracioso, Leticia Reis da Silveira, Maria da Graça Simões, Luzia Sigoli Fernandes Costa---------------------------------- 588

Information access in the digital era: document visualization strategy
Francisco Carlos Paletta, Armando Malheiro da Silva---------------------------------- 597

Integrating libraries, archives, museums and art galleries with Linked Data: initiatives study
Ana Carolina Simionato, Felipe Augusto Arakaki, Plácida Leopoldina Ventura Amorim da Costa Santos---------------------------------- 606

Knowledge Organization in a Web collaborative environment
Vânia Mara Alves Lima, Cibele de Araújo Camargo Marques dos Santos ---------------------------------- 613

Knowledge Organization Systems and cultural interoperability in open humanitarian settings
Quoc-Tan Tran---------------------------------- 624

Knowledge Organization systems used in European national libraries towards interoperability of the semantic Web
Dorota Siwecka ---------------------------------- 633

The Need to interoperate: structural comparison of and methodological guidance on mapping discipline-specific subject authority data to wikidata
Andreas Oskar Kempf---------------------------------- 644
Phenomenon-based vs. disciplinary classification: possibilities for evaluating and for mapping
Claudio Gnoli, Andreas Ledl, Ziyoun Park, Marcin Trzmielewski
653

Relationship status: libraries and linked data in Europe
Diane Rasmussen Pennington, Laura Cagnazzo
662

The Role of Knowledge Organization tools in open innovation platforms
Ricardo Eito-Brun
666

Subject access of Mexican television news broadcasts on the Web
Silvano Soto-Hernández, Catalina Naumis-Peña
674

Tags on healthcare information websites
Marit Kristine Ådland, Marianne Lykke
684

Towards integrated systems for KOS management, mapping, and access: Coli-conc and its collaborative computer-assisted KOS mapping tool Cocoda
Uma Balakrishnan, Jakob Voß, Dagobert Soergel
693

Towards the semantic annotation and the prevention of the loss of information of second opinion requests from rural Brazilian primary healthcare providers: the Q-codes use case – a work in progress
Melissa P. Resnick, Elena Cardillo, Marc Jamouille, Magdala de Araujo Novaes, Frank S. Shamene
702

Use of conceptual relations for semantic integration of scientific publications and research data
Luana Farias Sales, Luís Fernando Sayão
711

A Value-based approach to modelling interoperability in Knowledge Organization Systems
John Adetunji Adebisi, Babajide Samuel Afolabi, Bernard Ifesunor Akhigbe
720

VIAF and OpenCitations: cooperative work as a strategy for information organization in the linked data era
Erika Alves dos Santos, Marcos Luiz Mucheroni
729

Societal Challenges in Knowledge Organization

Author information for Knowledge Organization in the digital age
Wan-Chen Lee
739

Challenges of organization and retrieval of photographs on social networks on the Internet
Anna Carla Almeida Mariz, Raquel Oliveira Melo, Thales Almeida Mariz
746

Challenges to Knowledge Organization in the era of social controversies
Nathanaëla Andrianasolo, Adrian-Gabriel Chifu, Sébastien Fournier, Fidelia Ibekwe-SanJuan
754
Copyright infringement: between ethical use and legal use of information
Arthur Coelho Bezerra, Tatiana Sanches

Dealing with the paradoxes of customer opinion for effective decision support in churn management
Ayodeji O.J Ibitoye, Olufade F.W Onifade, Chika O Yinka-Banjo

Design Science as a methodology for the development of Knowledge Organization Systems in museological entities
Mariana Cantisani Padua, Maria José Vicentini Jorente, Natalia Nakano

Digital heritage: challenges and opportunities in the access and organisation of digital knowledge in contemporary societies
Tiago Trindade Cruz

Enhancement of digital heritage through digital social networks
Camille Rondot, Emmanuelle Chevry-Pébayle

Epistemology and Ethics of big data
Moisés Rockembach, Armando Malheiro da Silva

Information Design as knowledge technology in the organization of digital information environments on the Zika Virus and its effects
Maria José Vicentini Jorente

Intersectionality and the social construction of Knowledge Organization
Maria Aparecida Moura

Knowledge Organization in the digital age: the complexity of the global labor market
Francisco Carlos Paletta, Armando Malheiro da Silva

Memories in dispute, and reconfigurations of cultural heritage: for an Ethnography of museums
Samuel Ayobami Akinruli, Luana Carla Martins Campos Akinruli

The Mobile phone between challenge and expectations: a potential for information sharing between Algerian breeders and veterinarians
Radia Bernaoui, H. Peter Ohly, Kebbab Salim, Mohamed Hassoun

Negotiating participatory KO in crowdsourcing infrastructures
Ina-Maria Jansson

New ways to produce shared knowledge to improve cooperation in overcoming societal challenges in healthcare: the lever of innovative interface organizations in France
Christian Bourret
The Role of Neuroscience in information and knowledge appropriation
António José de Bastos Leite, Francisco Carlos Paletta, Maria Fernanda da Silva Martins, Teresa Silveira

Testing library catalog analysis as a bibliometric indicator for research evaluation in Social Sciences and Humanities
Maria Teresa Biagetti, Antonella Iacono, Antonella Trombone

What do museum website users expect from linked open data?
Alexandre Fortier, Elaine Ménard

What is happening about KOS in Spain: scientific production analysis, 2000-2017
José A. Moreiro-González, Virginia Ortiz-Repiso

Posters

Challenges in management and Knowledge Organization of documental heritage: key factors in the methodology of diligent search for orphan works
Rosario Arquero-Avilés, Brenda Siso-Calvo, Gonzalo Marco-Cuenca, Silvia Cobo-Serrano

A Connotative trust-based paradigm to minify societal challenge(s) in Knowledge Organization Systems
Ojo Stephen Aderibigbe, Bernard Ijesunor Akhigbe, Babajide Samuel Afolabi, Emmanuel Rotimi Adagunodo

The Contribution of research on information quality to Knowledge Organization
Priscila Basto Fagundes, Douglas Dyllon Jeronimo de Macedo, Enrique Muriel Torrado, Adilson Luiz Pinto

Creation of a domain ontology in CIDOC CRM OWL format using heterogeneous textual data related to industrial heritage
Eric Kergosien, Kaouthar Ben Smida, Rémi Cardon, Natalia Grabar, Mathilde Wybo

Diplomatic forensics: a necessary historical review for the analysis of the born-digital record
Juan Bernardo Montoya-Mogollón, Sonia Maria Troitiño

The French military documentary system to anticipate health risk: content and information classification
M.Tanti

Gamification as a system for developing knowledge in the classroom: a proposal based on an educational innovation project
Rosario Arquero-Avilés, Gonzalo Marco-Cuenca, Alicia Arias-Coello, Brenda Siso-Calvo, Silvia Cobo-Serrano, Carmen Soler-Vaquer
Interdisciplinary research organization: superimposing linked library data, linked research information and research data with interdisciplinary Knowledge Organization
Ingo Frank

Knowledge Organization and information retrieval in institutional repositories
Agnes Hajdu Barat

Knowledge Organization System interoperability: the cogitation of user interfaces for better interactivity
Nkechinyere Joy Olawuyi, Bernard Ijesunor Akhigbe, Babajide Samuel Afolabi

Mapping perspectival ambiguity in Bioethics: revisiting the viewpoint warrant
Denis Kos, Sonja Špiranec, Ante Čović

A Model for decision making (DM) in territory management: social implications
María J. López-Huertas, Diego Martín Oliva

State of the art of organization and administration of libraries in Brazil: preliminary results
Jaqueline Santos Barradas, Eliane Cristina Maceió Ferreira

Web archiving of elections and Brazilian possibilities
Moisés Rockembach, Lisiane Braga Ferreira

Workshops

The Politics of classification
Robert D. Montoya, Melanie Feinberg, Jonathan Furner, Gregory H. Leazer, Joseph T. Tenni

Thematic representation and metadata of photographic documents
Ana Cristina de Albuquerque, Ana Carolina Simionato

List of Contributors and Authors’ Index
Introduction

The digital and networked society in which we live is shaped by two paradigms that, since the middle of the last century, have been structurally changing our lives: the technological paradigm (Manuel Castells) and the paradigm of complexity (Edgar Morin). On the one hand, we are witnessing an increasingly accelerated evolution of technologies and communication networks, which means rapid change and constant dynamism; on the other hand, we live within a framework of complexity, challenges, new ideas and new forms of relationships and attitudes, which we cannot completely dominate given the fact that we are actors in the process and do not have enough critical distance to reflect on it. The information phenomenon has a central position in this context, since in symbiosis with technology it is the engine of change and innovation in which we are daily involved. In academia and in business, as well as in the professional and leisure activities that we pursue, we perceive that the world is changing in an accelerated way and that the processing, organization, dissemination and management of information in general requires study, research and development of theoretical and applied models that respond to the needs of society.

It is in this context that the 15th International ISKO Conference is held in Porto (Portugal) under the topic Challenges and opportunities for KO in the digital age. ISKO has been organizing biennial international conferences since 1990, in order to promote a space for debate among Knowledge Organization (KO) scholars and practitioners all over the world. The fourteen conferences already organized have focused on a wide variety of topics related to KO and its fundamental role in social, economic and cultural scenarios: tools for KO (Darmstadt, Germany, 1990), Cognitive paradigms in KO (Madras, India, 1992), KO and quality management (Copenhagen, Denmark, 1994), KO and change (Washington, USA, 1996), Structures and relations in KO (Lille, France, 1998), Dynamism and stability in KO (Toronto, Canada, 2000), Challenges in knowledge representation and organization for the 21st century (Granada, Spain, 2002), KO and the global information society (London, UK, 2004), KO for a global learning society (Vienna, Austria, 2006), Culture and identity in KO (Montréal, Canada, 2008), Paradigms and conceptual systems in KO (Rome, Italy, 2010), Categories, contexts and relations in KO (Mysore, India, 2012), KO in the 21st century - between historical patterns and future prospects (Krakow, Poland, 2014), KO
for a sustainable world (Rio de Janeiro, Brazil, 2014).

This time, organization of the International ISKO Conference is the responsibility of the ISKO Iberian Chapter, in association with the Faculty of Arts of the University of Porto, through CIC.digital - Center for Research in Communication, Information and Digital Culture - Porto.

The ISKO Iberian Chapter was founded more than two decades ago and, from 1993 to 2017, it has organized thirteen biennial conferences besides publishing the respective proceedings, which can be consulted from its website (http://www.iskoiberico.org/congresos/).

The topics under discussion in the 15th International ISKO Conference are intended to cover a wide range of issues that, in a very incisive way, constitute challenges, obstacles and questions in the field of KO, but also highlight ways and open innovative perspectives for this area in a world undergoing constant change, due to the digital revolution that unavoidably moulds our society. Accordingly, the three aggregating themes, chosen to fit the proposals for papers and posters to be submitted, are as follows: 1 – Foundations and methods for KO; 2 – Interoperability towards information access; 3 – Societal challenges in KO. In addition to these themes, the inaugural session includes a keynote speech by Prof. David Bawden of City University London, entitled Supporting truth and promoting understanding: knowledge organization and the curation of the infosphere.

In response to the call for papers, 162 proposals were submitted, including papers, posters and workshops, from which 110 papers and 24 posters were selected by the Programme Committee for oral presentation at the Conference as well as two workshops, the latter taking place after the conference, on 12th July. Although not all of the papers are included in the proceedings, the wide-ranging volume now published contains about 88% of the texts approved and it should be noted that some of these papers were also submitted to the Best Doctoral Student Paper Award established in recognition of PhD students’ work.

The Conference was supported by the University of Porto, the Foundation for Science and Technology - FCT, Banco Santander, and the company Recondicionados.pt, which sponsored the Best Doctoral Student Paper Award. We gratefully acknowledge all of them and express publicly our thanks.

At a time when we are constantly facing challenges and opportunities, we hope that the texts now published will be a useful tool for reflection and work, and constitute a further step forward in the field of KO.

Fernanda Ribeiro
Porto, May 2018
Keynote

David Bawden

Supporting truth and promoting understanding: Knowledge Organization and the curation of the infosphere

Abstract
This paper considers the response of knowledge organisation (KO) to a variety of problems and pathologies associated with the post-factual, or post-truth, society. It argues that there are no quick fixes, but that KO has several roles to play in mitigating these problems, particularly in the promotion of understanding, as well as the communication of information and the sharing of knowledge. Borrowing from Floridi's Philosophy of Information, it argues that KO, and more broadly library and information science (LIS), should address these problems as part of our role as 'curators of the infosphere'.

Introduction
This paper addresses two of the main themes of this conference, by considering a new foundational direction and purpose for knowledge organisation, as a response to certain societal challenges to the effective communication of information and knowledge. The new direction involves a realignment of purpose; from knowledge organisation being applied in the cause of the effective provision of information and documents to its application for the explicit purpose of promoting understanding. The societal challenges which this may address the the much-discussed problems of the post-factual or post-truth society, with its accompanying phenomena of fake news, the death of expertise, and the rest. This paper builds on a session at the ISKO UK 2017 annual meeting, devoted to these issues, and goes beyond it to consider how the promotion of understanding may in itself contribute to a solution. It uses Luciano Floridi's Philosophy of Information as a theoretical back-drop throughout.

Societal problems
The linked collection of social problems which have been described by such terms as 'fake news', 'alternate facts', 'post-truth society', 'post-factual society', 'death of expertise', 'filter bubbles', and 'social media echo chambers' are well known. Indeed, 'post-truth' was Oxford Dictionaries' Word of the Year for 2016, and 'fake news' was Collins Dictionaries equivalent for 2017, with 'echo-chamber' on its shortlist. Together, they describe a situation where objective factual truth is denied, expert informed, opinion is derided, and exposure to novel and challenging ideas is actively avoided. This situation, which is in many respects the antithesis of what library and information science (LIS) has sought to promote, and causes soul-searching within the theory and
practice of LIS (see, for example, Bawden 2017 and Cooke 2017). It has not arisen de novo: fake news has a long history (Cooper 2017), while the philosopher Bertrand Russell observed more that seventy years ago that "... most people go through life with a whole world of beliefs that have no sort of rational justification... People's opinions are mainly designed to make them feel comfortable: truth, for most people is a secondary consideration." (Russell 1942). Our present concerns are the culmination of a series of changes in the social and informational environment, brought into stark relief by political issues in Europe and North America from 2016, accompanied, and to an extent brought about, by a torrent of misinformation and disinformation (Clark 2017, Corner 2017, Wardle and Derakhshan 2017). A Pew Internet study carried out in early 2017 found a panel of experts almost evenly divided as to whether the problems could be ameliorated or would become worse over the next decade (Anderson 2017).

Naturally, many well-intentioned proposals have been advanced to remedy the situation. Some have addressed deep seated issues, in educational systems, in economic and social policy, in regulation of the media (including social media), and in political structures. Others, including several advanced from within the LIS community, have recommended more immediate and specific remedies; see Clark (2017) for an overview. Some have focused on the development of IT solutions, particularly with the algorithms used to filter news in social media and with automated fact checking and comparison (see, for example, Cooper 2017, Madrigal 2017 and Tomchak 2017). Others have advocated the enhancement of information and digital literacies (see, for example, Cooke 2017, Polizzi 2017 and Poole 2017), and of restoring the importance of expert objective fact checking and 'kite marks' (see, for example, Cooper 2017, Jirotka and Webb 2017 and O'Leary 2017).

My view is that, although these sort of initiatives may well have value, taken alone they will have relatively little impact. The problems and issues are deep-rooted, 'systemic' as Beckett (2017) puts it, and are not amenable to any 'quick fix'. As the philosopher of information Luciano Floridi emphasises, the more important the problem, the most it needs a long period of reflection to find the best solution. And as my colleague Lyn Robinson and I have suggested more specifically, LIS has no quick fix for these issues, and we should not pretend that we have; Beckett (2017) suggests the same for journalism in respect of fake news. We believe that LIS has a very considerable contribution to make, but it must be a deeper level than a tweak to an algorithm, a guide to information evaluation, or a reliance on the manipulation of big data (Robinson 2016, Bawden 2017, Poole 2017); as Floridi (2016) puts it, solving the problems of fake news and the rest requires a reshaping of the infosphere, our whole information environment and our interactions within it.

It seems to me that part of LIS's contribution to a longer-term approach to these issues will certainly lie in knowledge organisation. This has already been noted by
others. ISKO UK devoted a session at their 2017 annual conference to 'False narratives: developing a KO community response to post-truth issues' (ISKO 2017), and a plenary discussion of the Dublin Core Metadata Initiative in October 2017 debated 'A metadata community response to the post truth information age' (DCMI 2017). We will include points made at these sessions, and consider some other possibilities, later. First, we will make a slight detour, and think about the nature of understanding.

**Promoting understanding**

One way of expressing the problems of post-factual, expertise-less society is to say that it lacks a full and clear understanding of the issues facing it (Robinson 2016, Bawden 2017). Lyn Robinson and I have argued that LIS should take a new stance of focusing on the promotion of understanding as much as on the provision of information and the sharing of knowledge in an era when, for most people for the most part, information is provided through search engines, particularly Google, through a few encyclopedic websites, particularly Wikipedia, and through social media (Bawden and Robinson 2016A, 2016B). In this environment, we contend, the promotion of understanding falls, arguably uniquely, within the remit of LIS; this seems to be a novel suggestion, although it has been supported by Gorichanaz (2016, 2017), and fits within the Floridi-derived idea of LIS professions as 'curators of the infosphere' (Bawden and Robinson 2018).

There is, we may say, little understanding of understanding, in as much as it is defined very differently by various authors. We may note that Ackoff (1989) in his original formulation of the well-known data-information-knowledge-wisdom model, included understanding, which he characterised as 'an appreciation of why' as a high-level concept, between knowledge and wisdom. In the widely-used educational taxonomy due to Bloom, on the other hand, it comes as a rather low-level concept, above remembering, but below applying, analysing, etc. (Anderson, Krathwohl and Bloom 2001).

On the basis of an analysis of various conceptions of understanding (for details of which, see Bawden and Robinson 2016A, 2016B), we propose a definition of understanding, relevant to the purposes of LIS, following Ackoff, and situated within Floridi’s Philosophy of Information:

Information is taken to be well-formed, meaningful, truthful data. Knowledge is taken to be information organised in a network of account-giving inter-relations. Understanding occurs when a conscious entity, supported as necessary by information systems, appreciates the totality of a body of knowledge, including its interconnections. The extent to which the knowledge is incomplete, contradictory or false determines the degree to which understanding is less than complete.

Developing understanding in this sense would seem to be a worthy aim for LIS, and on which may go some way towards helping mitigate the societal problems noted above. However, we need to note that people may have an understanding of a topic, in
this sense, based on misinformation or disinformation, and may be impervious to contradictory information (see, for example, Requarth 2017), and this may be reinforced by emotional attachments to certain viewpoints (Beckett 2017, Poole 2017). We should therefore add a rider, to to effect that the extent to which someone is open to changing their views on the basis of new information, in effect to the extent of their curiosity, is also a measure of the completeness of their understanding. This could amount to a commitment (individual or societal), to accepting, indeed actively seeking, new knowledge even if it be potentially disruptive of current understanding (Bawden and Robinson 2016B). For LIS, this fits in well with the suggestions od Beckett (2017) that news media should provide content that is stimulating and challenging as well as relevant, and of Finch (2017) that libraries should be as much a safe place to indulge curiosity, rather than a trusted dispense of facts and information, or a repository of the truth.

Expressed in this way, we can see that the development of understanding may in itself be a powerful force for counteracting the problems discussed above; Bradley (2017) explicitly notes that helping people to understand and use items in their information environment is a role for libraries in countering the fake. It is important to note that developing understanding, in the sense meant here, is a broad and general approach, rather than a specific tool or technique, and goes far beyond didactic approaches to evaluation of information. Information systems are beginning to be developed to support understanding, from the relative conceptual simplicity of Google's Knowledge Graph, which integrates information from a variety of sources with the aim of giving a quick overview, to systems explicitly aiming a developing understanding from a corpus of sources. As an early example of the latter, see Donne et al. (2012).

We now turn to consider the ways in which knowledge organisation may contribute to these linked aims: the promotion of understanding, and the mitigation of the problems of the post-factual society.

**KO's contribution**

As suggested above, a variety of contributions to the amelioration of these problems have been suggested. At the risk of over-simplification, we can consider them under five headings.

First, we might notice the suggestion that an ontology, taxonomy, terminology, or glossary of the post-truth society, its pathologies, and potential solutions, may be of value in itself, as a way of clarifying the concepts and their inter-relations, and as a guide to action (Bradley 2017, Clark 2017, Poole 2017, Wardle and Derakhshan 2017). For the most detailed example yet extant, see Synaptica's *Post-Truth Forum Knowledge Base* at https://www.posttruthforum.org.

Second, there is what we might think of the classic, if limited, response of KO:
adaption of methods of resource description, and revision of existing descriptions. One popular example of the former is the idea of 'credibility metadata', the addition of terms aimed at reducing misinformation and disinformation by establishing veracity of resources (Wardle and Derakhshan 2017). This may involve markers of location, time, etc. on items, or metadata to note 'quality factors', such as that a source has a corrections policy, or that the author of an item has written on the topic before (Cuellar 2017). Conversely, indexing may directly address the 'fake' nature of an item, as in the idea of adding a term for 'satirical article' (Quinn 2016, Cuellar 2017). The latter, by which discredited materials may be identified as such, is well exemplified by the controversy over the reclassification of Holocaust denial literature as 'historiography' rather than 'history' (Simon XIX 2017).

These responses are generally implemented by traditional intellectual metadata construction. The third category is the use of automated classification and indexing to attempt to identify and categorise fake news and other pathologies of the post-factual society; of the many developments of this kind, a good, albeit simple, example of a classifier to distinguish genuine news items from fake is given by McIntire (2017). More complex examples, based on more sophisticated machine learning techniques and classification techniques are likely to play an increasing role (Cooper 2017).

Fourthly, there are those KO techniques which directly support curiosity which, as noted above, is a powerful force for finding alternative perspectives, breaking filter bubbles, and building understanding. In this respect, another long-established aspect of KO, classification techniques with their ability to show both hierarchical and associative relations, may be of particular importance.

It is worth considering, from the perspective of the issues discussed here, whether any particular form, or theory, or classification may be most appropriate. In particular, pragmatic or critical classification (Hjørland 2017) appears to be something of a two-edged sword. On the one hand, it may support, or reflect, an understanding of the world helpful to an individual or a group in developing understanding, and coincide with their emotional responses to issues; on the other hand, such an organisation may simply reinforce filter bubbles. It may be that systems could be developed to allow a ready comparison of alternative classifications, assisting curiosity-driven explorations of different perspectives. We may also need to consider the status of classical classifications, based on a single agreed picture of the world, and approximating as closely to truth as may be possible. Are these sustainable, at a time when alternative facts seem as viable as any other, and when expertise is said to be said? They appear to be the antithesis of this negative viewpoint. The status and role of classification in the post-truth era seems to be an area in need of thoughtful research.

Fifthly, and finally, it seems to me that to deal adequately with current problems, KO must fully recognise the deep and irreversible changes in the information
environment brought about by the shift to what Floridi categorises as 'infosphere' and 'onlife' in which our digital and physical lives merge, and information, contextual and mobile, is central to our society, and indeed to our humanity (Floridi 2014). This is a long-term and far-reaching challenge, encompassing and going beyond the challenges of the post-truth society, and one in which KO should have a unique position.

Conclusions
There are no quick fixes to the problems set out above. KO cannot solve these problems alone, any more than can the wider LIS discipline; more far-reaching and structural changes, particularly educational, are needed for that. However, KO can play a significant role in improving the situation, using a combination of classic KO concepts and familiar KO practice, integrated with newer technological and organisational environments. In this way, by opposing misinformation and disinformation and promoting understanding, we may justify a claim to be curators of the infosphere.

Acknowledgements
I am grateful to Nick Poole and to Vanda Broughton for helpful comments and advice.

References


Foundations and Methods for Knowledge Organization
Among dictionaries, reference works and concept systems: can the terminology in International Standard ISO 5127:2017 contribute to Knowledge Organization?

Abstract
The purpose of this paper is to introduce the main features of the new International Standard ISO 5127:2017 “Foundation and vocabulary of information and documentation” and to examine whether, and how, it could contribute to current knowledge organization (KO). Equally, this paper contains some considerations with regard to commonalities between KO and “terminology” and to principles and structure of both types of systems. ISO 5127 is, at the same time, a companion document for ISO/TC 46 “Documentation”, and thus also serves as its main reference document, systematically bringing together the terminology developed, required and used for the work undertaken by that committee.

The task of ISO 5127:2017 is
- to provide basic components for a descriptive outline of the information sector (field) and the procedures with which society handles predominantly scientific information work and its information processes;
- thus, also to contribute to KO systems and procedures.

In some of its functions “terminology” is a meta-tool, an instrument that helps in the creation, and maintenance, of other systems such as a KO system.

This paper is not so much a study of epistemology or the general foundations of KO systems as a look at how the domain “information field” is reflected in the concept system of a “terminology”, how it is structured and how this terminology relates to (other) KO systems.

The need for “terminology”
An arbitrarily picked example demonstrates directly the need for terminology and definitions, while at the same time indicating how problematic this area is:

“As far as possible the terms used in the model to identify entities or objects, their attributes, and their relationships have been derived directly from the code itself. The definitions for the terms used to identify entities, etc. have also been derived, to the extent possible, either from the glossary or from the wording of the rules themselves, in some cases in a modified form. Where definitions could not be derived from the code itself, they have been derived from other sources such as the ALA Glossary of Library and Information Science, and the Functional Requirements for Bibliographic Records. In other instances definitions have been developed specifically for the purposes of the model.” (The Logical Structure … 1999, p. 2)

Right at the beginning of their ground-breaking report on new (and international) cataloguing rules, the renowned information expert Tom Delsey and his colleagues remind us of the fundamental need for conceptual clarity, and spelt-out definitions, in any undertaking in the information field and in KO. Only on that basis is it possible to conceive projects, carry out tasks and ensure some mutual understanding of people working together on an issue. At the same time, it is clear that “understanding” and
“precision” vary widely: in practical procedures definitions may often be implied and intuitively understood by those involved; they may be only loosely worded (simply taking the form of examples, enumerations, demonstrations, etc.); they may come down to a mere approximate circumscription; or, in practical reality, they may just be conceived ad lib. It increasingly seems to be the case today that despite the continuous need for very painstakingly precise definitions (in the very nature of programming languages, for example) like those of IT functions, at the same time, even projects, seem to rely to a larger degree on intuitive understanding and its development rather than on the need to have very precise definitions at every occasion. On the other hand, however, procedures and requirements may be worded with the utmost precision and be so perfectly fitted into a concept system that they achieve the highest rating on a “terminology acceptance” scale. The more technically complex an issue, and the more it requires a deeper understanding (including scientific research), the greater the need may be felt for a more formalized, precisely delineated understanding – that is, for a “definition”. A good example of that is the construction of “data elements” as basic units for processing information semantically and technically, which strongly rely on high precision definitions (see also ISO 8459:2009). Therefore, what constitutes our “terminology” project and procedure, as described in this paper, and how does it fit into a whole landscape? Where are definitions when we need them and how can we find them? Do we have to rephrase them again each time, with every new instance of a constantly recurring issue?

This begs that question, why not simply pick out a definition from the relatively large number of existing vocabulary reference works (the ISO 5127:2017 bibliography lists about 150), such as Harrod’s Librarian’s Glossary, the Glossary from AACR (or RDA) and ISAD, British Standard DD 247 (withdrawn in 2012), or one of the many others? The answer is because they do not often conform to essential definition requirements, and, looking at the information field overall, when used concurrently and side-by-side, they often do not match with one another.

Although, on the one hand, there has been an “explosion” of ever-varying new developments, with a wide range of differing projects and actions, there has, on the other hand, also been a growing trend over recent years to build more extended, formally precise and coherent systems like those of “terminologies” – intended for the realm of concepts and intellectual clarity, but now also for numerous other practical applications. The most striking example in the cultural and thus the documentation sector is the Conceptual Reference Model (CRM), which has emerged from ICOM-CIDOC as one of the now many “(domain) ontologies”, expressing strictly logical-based entities (classes and properties) and even inverse relations with the character. ISO 21127:2014, with some 120 classes and 80 properties assigned to it, is a model example of a concise, and partly comprehensive terminology within a specific subdomain in the “information field”. Its transformation into FRBRoo and into archival systems shows how far such
systematic endeavors have now gained ground in the information field as a whole.

“Knowledge Organization” and “Terminology”

“Knowledge Organization” (KO) and “Terminology” might seem like natural twins. Although they are as yet often carried out in quite different frameworks and organizations – mostly independent of each other – they share a clear common basis in being grounded in identifying and ordering ‘concepts’. For founders/prominent representatives of KO (like Ingentraut Dahlberg), pioneering and decisive work on concept theory took place alongside equally fundamental work on classification systems and, later, thesauri, each line of research being firmly based on the study of concepts. “Terminology” (concept and term systems) expresses “knowledge” – just as “knowledge” can hardly ever be conceived without “terminology” (clarification of meanings) being integral to it.

It should thus come as no surprise that, as stipulated by terminology theory and by the relevant standards ISO 704 and ISO 10 241, in each technically specified field of human knowledge and practice, specialized terms should be clarified, and their meanings made obvious – especially for the starting point of translation(s). Accordingly, every such field would have an explicitly verbalized terminology at its basis (possibly fixed in an ISO standard itself). Hence, the field of “Documentation” / KO adheres to ISO Standard ISO 5127.

ISO 5127 terminology in its structuring of the information field

ISO 5127:2017 is intended as a vocabulary that encompasses the whole information field, with all its sectors and professions – much as universal classifications also do so. In this regard, ISO 5127:2017 is compliant with other advanced approaches such as setting unified procedures to describe all types of information material in various sectors, as with the new international cataloguing code “RDA Resource Description and Access”. ISO 5127:2017 contains 2,000 definitions and their terms, arranged into 18 sections and 131 subsections.

ISO 5127:2017 has broadened its coverage in comparison to its previous version (ISO 5127:2001 containing 1,200 definitions). This comes in response to its objective of better interconnecting the various branches in the information sector, such as archives, documentation centres, photo or phono archives, museums, libraries, publishing, scientific editing, documentology, and so on. ISO 5127:2017 aims to reach out over and across the various subdomains and, with this approach, also intends to foster interoperability and interworking. Both these aims are in perfect harmony with a central purpose of large classification systems. ISO 5127:2017 covers the ever-growing number of roles played in the current information field by, firstly, various kinds of communication (19 definitions, including “3.1.8.13 oral communication”, for example) and, secondly, by differentiated forms of knowledge (15 definitions, including “3.1.1.32
tacit knowledge”, for instance) This is done, thirdly, with due consideration given to the
components of IT and to digital procedures which are at the core of today’s processing
in KO (even though “traditional” analogue practices continue to play an important role
and need still to be taught and maintained, given the preponderance of ‘traditional’
analogue documentation techniques in many areas and in many countries of the world).

A comprehensive terminology like ISO 5127:2017 should supply the essential
terms/concepts required to understand the information field both as a whole and in terms
of the various components that constitute it – just as every classification for the same
subject would and should. In addition, it may list certain special entries that clarify little
known or obscure specialist terms in the field (such as “3.5.1.07 calendar (1)”). In
addition, a proper terminology also to some degree preserves branch-specific vocabulary
where this continues to be needed (despite the overall confluence of the branches within
the information sector). This may result in a more detailed breakdown for some
subdomains; as in occasional provision of the sometimes variant terms used in differing
domains for substantially the same thing (e.g. “3.4.5.01 abstract” / “3.4.5.02 brief”); or
in showing concepts (and terms) that are only used in one specialist sector (e.g. archives
vs. libraries vs. media documentation, etc.), such as “3.5.4.07 administrative history”.
All these remain very well respected and yet, at the same time, have become part of the
general context. Further examples may be the very different meanings of the term
“collection” in a library, an archive, or a museum, or the careful differentiation of
different “values” which has been established over time associated with archival
documents (3.3.1.01 / 09) – not being so relevant, since it addresses a different level of
document study, for libraries, but usefully applicable in the similar work of museums.
Thus, ISO 5127:2017 terminology drafts and structures also by revealing details and
differentiations rather than simply being one-sidedly driven by the influence of the
technological basis in the information sector, which is now rapidly changing and
evolving every day.

Consequently, with respect to all its features, ISO 5127:2017 should also come to be
regarded as a useful and systematic teaching tool in the information professions.

Terminology and KO as mutually related in their historical development

KO systems (e.g. classifications, thesauri, etc.) can be understood as consisting of at
least two layers: the contents which they actually express (such as “metallurgy, mining,
chemistry, furniture production, agriculture, education”, etc.), and the underlying
entities and construction principles that are inherent to KO systems, on which they are
built (entities and techniques of “knowledge organization”/proper documentation, such
as “class”, “indexing”, “broader term”, “hierarchy”, “caption”, etc.). A third layer is thus
the representation of this latter content itself as content within and via the KO system,
as a technical subject in and through the system (e.g. class 025.4 in UDC). It is now
interesting to see that attempts at “definitions” in the strict sense of the word, more
“formal” (explicit), more conceptually precise definitions – not only of those technical subjects (like “metallurgy...”) reflected in the classification systems, but of some basic entities that constitute (or are an instrument for) the KO/classification system itself (meta-level) – at the beginning of the classification work, did not by any means stand for DDC, for example, or for drafting the first UDC version(s), which began as of 1892. Admittedly, the first “formal” setting out of definitions (= precise verbal expression of concepts) – for cataloguing rules, in this case – is encountered as early as 1905 in the work of the IIB. But it was not before 1928/1932/1935 that Otlet and then a few of his colleagues set out a small initial set of 14 basic definitions (e.g. “document”, “documentatoir”, “rubriques”, “divisions principales”, etc.) (Ermert 2008:372-374). Moreover, precisely at this time, when such new endeavours coincided with the newly emerging theory and application of terminology outlined by Eugen Wüster (1932) in his ground-breaking book, another distinct area of study related to KO emerged, called “Terminology Research/Science”. For the first time, this study very explicitly stated the difference and relationship between “concept” and “term” (and linguistic forms of the latter). It also stated that – in technical and scientific applications – no term should be applied without proper clarification of its meaning/definition, and that, in fact, all scientific terms should be concept-based: concepts should be defined and presented in relation to each other first, and only then – after an understanding has been gained of their interrelations, their possible overlaps and the delineation of each individual concept – should the appropriate term(s) be allocated to them. Hence, this practice of “terminology”, part art, part science, can from thereon be recognized as adding to the emerging idea of KO in general, as a third cornerstone (alongside initial “classification” and “indexing”).

The overall distribution of the top leading ISO 5127 groupings in the information field

KO, and its terminology, can be understood in a narrower and a broader sense (the latter properly being called the “information field”). It may be said that “documentation” / “knowledge organization” seems to have initially emerged consisting of an “inner circle” of core features, namely:
- classification (later extended by “thesauri”)
- indexing, for the purpose of finding documents (“resources”) by their contents,
- supplementing from earlier terminology attempts/projects

Since then, our view on “documentation” / “knowledge organization” has broadened considerably into what it is today. Thus, ISO 5127:2017 is currently organized to reflect six broad areas of the information field, stemming from an analysis of “information” and its resulting systematic interrelations, as shown in the chart below:
Even one of the original core groups (“classification”), from the “inner circle”, already invites an important question for a terminology like ISO 5127: which sources should be drawn upon? For “thesauri”, ISO 5127 draws quite extensively on the elaborated vocabulary in the TC 46 thesaurus standard ISO 25964-1/2. For “classification” such a model turns less to pre-existing material; ISO 5127:2017 presently offers 34 definitions on this subject. Certainly, it should draw heavily on both recent and established textbooks on classification such as that of Vanda Broughton (2015). And here – as elsewhere – it would perhaps also be useful and necessary to reproduce some of the systematic vocabulary established by the early founders of KO like (Wüster or, in this case:) Ranganathan and which is shown in Indian Standard IS 2550, among others. Besides that, a look into the carefully elaborated Russian “Vocabulary of UDC terms” (1986) might be rewarding.

Based on its foundational concepts, as shown below, (3.1.1.01 object, 3.1.1.02 concept, 3.1.1.15 data, 3.1.1.16 information, 3.8.5.03 class) and on the grouping of the material along the broad areas listed in the graph above, ISO 5127 then goes into further detail with a series of main subgroups, logically emanating from these primary divisions, as follows:

- 3.1.6 Writing, script systems
- 3.1.11 Basic operations on data
- 3.3.3, 3.4, 3.5, 3.6.1, 3.6.3 extensive lists of document types (e.g. “primary documents” and “tertiary documents” (reference works, catalogues, etc.) / aggregations of documents (collections)
- 3.2, 3.7, 3.8 Documentation / KO operations (description, cataloguing, classification, indexing, mining, etc.) and tools (classifications, thesauri, indexes, discographies, etc.)
- 3.2.3, 3.2.4, 3.6.4 Institutions and professional roles in documentation / KO
- 3.13.2 / 3.13.5 Intellectual property and data protection.

At the very least, an extensive treatment of document types (308 + 65 definitions in ISO 5127:2017) seems very recommendable here – identifying them has been an important objective of universal classifications from the beginning. This would enlarge knowledge about how information materializes and is transported. It has also played a considerable role in characterizing for users the form in which they encounter information signalled by documentation services, and in characterizing it in bibliographic services provided by libraries. Identification and coding of document types has also been urged several times by KO experts for the development and enlargement of the UDC (Ermert 2016).

A terminology built on foundations

Terminology standard ISO 5127 was initiated between 1967 and 1969, at the same time as the 36-part ISO 2382, and, like its “big sister”, is carefully based upon general key terms (data – information – data processing – automation – data medium - signal – numeric – digital – analog – hardware – computer, etc.). Following its launch, ISO 5127 then formed its basic foundational groupings around certain pairs of fundamental concepts. They serve as starting points and underlie all subsequent entries in the concept system as a whole. (A similar procedure is followed in KO systems like classifications.) It is vital that they invariably form pairs (or, more precisely, triads or quadrigas), and need to remain interwoven throughout the system. These intertwined concepts should be represented and defined as consistently dependent on each other, as forming logical and inseparable conjunctions that constitute the whole. The concept system, and thus the information field represented within it, depends directly on them. These are the pairs that constitute the foundation for all further entries in ISO 5127:2017 which follow subsequently:
- (signal), data, information, knowledge, relation, ontology
- concept, class, facet, category, definition, designation, term
- communication, medium, data medium/data carrier, document
- types of relations between entities (hierarchical, associative, etc.), order
accompanied by a bundle of general concepts such as “3.1.1.13 system”, “3.1.13.27 entity”, “3.1.13.28 instance”, “3.1.1.09 set”, an “network”, which are required for any systems analysis and systems building as well as in programming, data modelling and IT systems design.

The basic groupings and concepts in the ISO 5127:2017 terminology also respect the fact that the information field these days is constituted by not one, but a multitude of
media of very different natures: not just the printed text document as in former days, but also audio media, film, artistic documents, radio and TV, digital media, Internet, among others. ISO 5127:2017 incorporates part of all this in appropriate subchapters.

**Structure and principles of “terminology” in clear correlation with KO systems**

Like classifications, KO systems traditionally consist of four main components (thesauri playing an intermediate role):
- a class name/caption/class heading (corresponding to the term in a “terminology”), which represents a concept, more or less carefully chosen and determined
- it is usually assumed that this class name is clear, and its meaning intuitively understood. Normally, no explicit definition of the concept underlying the class (name) is given.
- a notation preceding the class name and showing the place of this entry within the entire system
- the place of the class within the system as a whole is indicated only by the notation. This place is not usually obvious from the class name.

The close similarity between a classification and a “terminology” becomes clearly apparent from the following, quite similar overview of the respective features of a terminology:
- a term (corresponding to the class name in a “classification”), which represents a concept, more or less carefully chosen and determined
- an explicit formulation (definition) of the concept underlying/expressed by the term
- the place of the term (and thus, the concept) within the entire system becomes directly apparent from the definition of the term.
- (sometimes, numbering or even notation precedes the term/definition entry and indicates its place within the system.)

<table>
<thead>
<tr>
<th>“Terminology” (e.g., ISO 5127)</th>
<th>Classification (e.g., UDC, DDC, ..)</th>
</tr>
</thead>
<tbody>
<tr>
<td>term</td>
<td>class name / caption / class heading</td>
</tr>
<tr>
<td>definition</td>
<td>-</td>
</tr>
<tr>
<td>(- -)</td>
<td>Notation</td>
</tr>
</tbody>
</table>

Exactly parallel to classification systems determining the class names to be included in their systems, ISO 5127 selects terms, defines their concepts, both for entities of an abstract nature (e.g. “3.2.1.20 ofness”, or actions, e.g. “3.2.1.26 identification”) and those of a concrete nature, i.e. which directly correspond to concrete real-world entities (e.g. “3.4.7.48 score”, or “3.4.7.42 letter (2)”).
The principles of terminology that enhance it as a part of knowledge organization

The building and placing of the concepts within a concept system considerably adds to their understanding and determination; showing their place in a system adds ‘knowledge’ even to a given and already defined entry in the terminology, beyond even the existing explicit verbal expression in the definition itself. This, in turn, highlights the important contribution of certain KO systems to “terminology”, the importance of classification for and within a given concept field and thus simply for it being a pure mass of definitions. It also marks a key difference between strict terminologies versus alphabetically sequenced dictionaries (even if they apply – as often they do not – the rule of starting each individual dictionary entry with the relevant broader concept/term). In contrast to this – to quote another example – a well-structured and carefully elaborated thesaurus like the AAT Art and Architecture Thesaurus very well shows the system, with its hierarchies, by/through its thesaurus display; but it does not at all make them systematically apparent through the explanation in the individual thesaurus entry itself. These are merely explanations, often lacking a coherent definition style, and are quite rightly called “scope notes” rather than “definitions”. Though these “scope notes” clearly expand the explicative capacity of thesaurus terms, they are still not subsumed into a system through the wordings themselves.

Definitions can be formed along different criteria, and an item to be defined can present more than one relevant property/characteristic. Hence, several different concepts are often present in one item/entity – as in the case of compound subject content (ISO 5127:2017 “3.8.5.10 superimposed class”, “3.8.5.11 composite class”). Consequently, these point to different broader concepts/terms and thus to different places in the total system. (Therefore, for one unified, integrated terminology there needs to be some harmonization regarding how and when these entries shall fit into one overall concept system). This problem - well known within classification theory as a whole – is not easy to solve in the linear presentation of a terminology or any KO system on paper (unless the system is completely dissolved into a fully faceted one – but even then, the question remains of the filing order for showing the composite notations that have been created). In this respect, the UDC has long included the special feature of “Common Auxiliary Tables”. British Standard DD 247 (1998) – a terminology – gives the definitions/entries in alphabetical (linear) order, adding to each one all the classification notations it requires, resulting from the application of a special classification to the terminology stock. Thus, the table presented in DD 247 is the alphabetical (definition) display, rather than that of a classified display. For example:
If there is indeed – as indicated in the introduction (“The need for terminology”) – a ranking in “terminology acceptability” that ranges from very loosely formulated, through restricted and sometimes circular/contradicting, to very carefully presented and formally elaborated terminology as its highest rank, then this highest rank certainly befits clearly the most logical and strict terminology requirements as put forward by ISO 704/ISO 5127. According to them, three features are definite minimum requirements: a) there must be no circularity between the individual definitions in the system; b) every definition must refer to / be built upon the essential property of the object and must be limited strictly only to between one and three properties/characteristics that clearly distinguish this object from other, neighbouring ones; c) each individual entry of a terminology should, as far as possible, be supplemented by its antonym(s). d) open enumerations and opening phrasings such as “General term used for ...”; “Any of ...”, “normally understood to include …”, and the like, should be avoided. Terminology is thus essentially quite different from a (mere) dictionary – whose systematic conceptual presentation is necessarily weaker but which – due to its nature – sometimes provides a wider range of useful side information on an entity, such as grammatical information, particular features, history, specialties, etc. It also follows from the strict requirements of ISO704/ ISO 5127:2017 that, despite careful collection and assembly of terminology from ISO/TC 46 (and other ISO) standards, ISO 5127:2017 would (need to) clean, revise and edit quite a bit of this terminology, since it often does not live up to the strict requirements set out by ISO 704 / ISO 5127 rules: Terminology in individual standards or when confined to limited projects is often too narrow (e.g. misplaced concreteness; impermissible restriction to one single application case only; open enumerations in the definition; circularity, etc.), with, for instance, terms unrelated to each other. Examples: “User – recipient of library services”; “visit – person entering the library premises”; “electronic service – library service delivered via electronic means”; “opening hours – hours at which a library is accessible for users”.

**Extent of coverage of core “information field” terminology in ISO 5127:2017 and in some KO systems**

ISO 5127:2017 and major classification / KO systems, to a great extent, cover the same ground: much as does ISO 5127:2017 and major KO systems like UDC, DDC (in their respective parts), the ASIS&T thesaurus, etc. This clearly delineated scope covers
exactly the same matter in both types of systems – “terminologies” and KO/classifications extending to the information field. Therefore, one could ask whether they would not complement each other (identify and add missing entries in one system from the other, or supply required clarification of concepts already included in one system from the other)?

The interrelations and overlaps are obvious, so that these endeavours could, I believe, take advantage of each other. If ISO 5127:2017 is balanced and comprehensive enough in its coverage of the information field, it could thus also be one source for possible underpinning and enhancing of (other) KO systems. The KO systems (classifications, thesauri, etc.) on the other hand would provide tentative overviews of information field-relevant entities, show their grouping and presentation as well as their interrelations in a classificatory manner. “Vocabularium bibliothecarii” (1962) interestingly enough had arranged the sequence of the term entries along the UDC number. It had proven to be such a great advantage that (unfortunately, only during a certain period of the 1970s and 1980s) two experts from the UDC community took part in the work on ISO 5127. However, it is noteworthy to mention in this context that – possibly depending upon several different viewpoints being relevant for subdivision – what is attempted to be shown as one consecutive stream of the information field in ISO 5127:2017, is in some major classification systems split into several separate main groups (e.g. archive issues, artistic documents, photography, printing, and others in very different places). Their interrelation is thus not that fluent as might be desired – when the issue of “information” is the pervading perspective that structures the entire classification (presently, this is the concern of sciences or, rather, scientific domains).

Table 4: Coverage of core “information field” terminology in some KO sources; x x: “approximate match”

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x 3.5.1.26 x</td>
<td>(017.092) / 025.341.1</td>
<td>accession number</td>
<td>-</td>
<td>x accessioning x</td>
</tr>
<tr>
<td>(ex contained ?)</td>
<td>x 1d(048.83 x</td>
<td>analytic-synthetic classif.</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>caption [1 = class heading]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3.5.8.29 caption [2 = text under an illustr.]</td>
<td>7.024.4</td>
<td>-</td>
<td>3.4</td>
<td>-</td>
</tr>
<tr>
<td>3.5.8.03</td>
<td>517.518.2 / 57.063.2</td>
<td>class</td>
<td>3.10</td>
<td>x classification</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>broad classification</td>
<td>-</td>
<td>x classification</td>
</tr>
<tr>
<td>3.1.1.15 data</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>data</td>
</tr>
<tr>
<td>3.1.1.16 inform.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>information</td>
</tr>
<tr>
<td>3.1.1.17 knowl.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>knowledge</td>
</tr>
<tr>
<td>3.4.4.01 file (2) / 3.1.13.16 record</td>
<td>35.077.7</td>
<td>-</td>
<td>-</td>
<td>records</td>
</tr>
</tbody>
</table>
Table 4 shows in non-representative examples how core parts of the information field are represented in some KO systems and how these might possibly merge or complement each other (e.g. in filling gaps).

For a “terminology” in the proper sense – as outlined in this paper – it is inviting and also vital to translate the once established concept system (herein known as the “information field”) into many natural languages – making the system operational and applicable by transcending borders and overcoming language barriers: for ISO 5127 just as much as for KO systems, namely classification systems, thesauri, etc. UDC and DDC with their many national language versions are “unbeatable” in this respect. But it is very gratifying to see that the precursor edition ISO 5127:2001 has been translated into 16 languages, with 5 additional national editions. Among them there is, for example, a full Spanish, Polish, Chinese or Vietnamese language version. ISO 5127:2017 thus aims to be an ambitious and legitimate successor of such widespread classical tools as the UNESCO “Terminology of documentation” (1976) (5 languages) or the “Vocabularium bibliothecarii” (1962 / 1967, 1968, 1972) (12 languages).

To sum up: with its several features outlined in this article, ISO 5127:2017 indeed performs a three-fold function, fulfilling each to a varying extent: a) it enables a systematic overview of the information field by providing its basic concept system; b) by first supplying essential information on each entry and on the system as a whole, it qualifies as a reference work on the information field and for those active in it; c) by listing core terms (and, in future, foreign equivalents) of the information field, it is a source of initial information and assurance when only seeking a specific term or – for many possible purposes - simply a good insight into the vocabulary (word stock) of the information field.

**A published standard and its immediate revision**

ISO 5127 considers itself as a kind of a bridge. It may not be the absolute foremost reflection of the most recent inventions and developments in, say, IT, artificial intelligence, programming and internet design techniques, yet it aims to assemble those entries and that knowledge which can by now be considered foundational and effectively secure for any KO work today. This includes a proper part of ISO 5127:2017 which at first glance may still look “conventional” (rather than “digital”) as in classic libraries, publishing and document processing. However, (since the underlying concepts remain true, regardless whether they are in analogue or digital form – a “book / monograph”
remains a “book / monograph”) this good part is not only necessary for continued future understanding, but also for operating, all these items being in some new (digital) form. It also keeps knowledge about “traditional” ways of KO and its systems available which still forms a prominent part of the world’s information sector and activities, both today and into the future.

If not acquired as a self-contained document (now available at a reasonable price), ISO 5127:2017 or individual entries from it can also be obtained through accessing www.iso.org/obp.

Apart from all this, ISO 5127:2017 is perhaps a bit special, with regard to its functioning and nature as a (formal) standard: though the product of a long research process and a lengthy experimental phase, it has already started to undergo revision in 2018, directly following its publication in 2017. This somewhat unusual procedure is due to the fact that, in the relatively short time-period conceded for the work on the 2017 edition, not all items could yet be solved to a satisfactory degree. A more logical sucession of the entries in the beginning needs to replace the somewhat messy one which is presently there. A number of entries need to be added to bring the document finally up to a more or less satisfactory level covering the main basic components in the information field as of the decade beginning 2010. French (and possibly German, and Spanish or, questionably, Polish) terms need to be added to the concept entries. This immediate revision may also be helpful to eliminate mistakes that have been identified until now, and to integrate some of the feedback the standard will hopefully receive. However, interest in terminology and definition work somehow does, indeed, seem to follow up-and-down cycles, and the cycle we are presently in does not appear to be the liveliest time for interest in such work and tools.

ISO 5127 is an attempt at a big “leap forward”. Its presentation of a collection of material in its current state needs to be refined, improved, and tested against theory and practice. It calls on the professional community for collaboration and looks to continuously improve the shape of terminology as well as potentially contribute to further enhancement of KO systems.

References


ISO 25964:2011/2013 Parts 1 and 2 Information and documentation – Thesauri and interoperability with other vocabularies [includes classifications]. Geneva: ISO.


Natalia Bolfarini Tognoli, Ana Célia Rodrigues

An Analysis of the theoretical and practical application of Diplomatics to archival description in Knowledge Organization

Abstract
The present study aims to elucidate the relation between archival science and diplomatics based on how the latter may be applied to archival description, encompassing the ideas of identification, representation, and organization. We thus introduce the concept of archival knowledge in the context of the methodology of identification, an intellectual task whose object of study is the produced entity and its records, with a recognition process based on the parameters of diplomatics.

Introduction
Archives are the necessary and obligatory product for the functioning of any organized society. Every activity performed by an institution or individual uses records to enforce or affirm a particular action. Records are created or received by an entity or physical person in relation to their activities (classic archival definition), thus serving to ensure accountability, especially in a democratic society with an obligation to take account of events, both from an administrative and historical point of view.

Over recent centuries, records, archival institutions, and archival science have evolved considerably in every aspect, as a consequence of the accumulation of increasingly compact and broad-based relations, leading to an exponential expansion in the requirements and uses of records necessary to act, negotiate and live (Delmas 2010).

Archives can be conceived as information systems, alongside libraries. Accordingly, both institutions share the purpose of organizing specific knowledge produced and recorded by society through its public and private institutions, in order to enable its portability in space and permanence in time (Smit and Barreto 2002) and, as a consequence, to promote its comprehensive accessibility (Guimarães and Tognoli 2015).

Archival Science is an independent field, but knowledge organization of archives should also be considered a part of KO (Hjørland 2016). In this context, we believe the discipline can benefit from research conducted in knowledge organization (KO) and its approaches, since KO involves the describing, representing, filing and organizing documents and document representations, as well as subjects and concepts, as performed by both humans and computer programs (Hjørland 2008; Tognoli et al. 2013).

As pointed out by Smiraglia (2014, 31), “archives and records repositories, like libraries, serve a critical role in their social milieus, that of preserving and disseminating...
the collective knowledge of their cultures.”

But what is this collective knowledge? What do we understand as knowledge in archival science and how do we organize it?

In the last thirty years, archival science has been invited to rethink its principles and methodologies, in order to deal with the production, organization and use of records in the digital environment. In such a specific context, currents of thought have emerged in different countries, such as postmodern and integrated approaches in Canada, post-custodial archival science in Portugal, the records continuum approach in Australia, and archival diplomatics in Italy.

Research in the latter has supported the methodologies of archival knowledge organization since the 1960s when the object of study of archival science was first identified with that of diplomatics, namely records (Bautier 1961).

In the late 1980s, diplomatics was revisited by archival science and found a space within the discussion of the theoretical construction of archival identification for reflection on research into documentary genesis and its relevance to the standardization of record management and technical treatment of records accumulated in archives (Rodrigues 2016).

The present study aims to elucidate the relation between archival science and diplomatics based on research into archival description – understood as a process of archival knowledge representation – and on its relationship with archival identification, "a research methodology that is developed, in the parameters of scientific rigor, as a preliminary and necessary task to the archival functions such as classification/arrangement, appraisal, description and also for the planning of documentary production" (Rodrigues 2012), contributing to the construction of what we call archival knowledge.

Finally, it presents some results of postdoctoral research developed at the Fluminense Federal University - UFF, in Brazil, whose goal was to construct theoretical and methodological elements to characterize diplomatics and its method as necessary for the organization and representation of archival knowledge, within the scope of two main archival functions, classification and description, and its relationship with archival identification.

**Diplomatics and Archival Science: two disciplines, one object**

The beginning of archival theory (Schmidt 2015) came with the promotion of the principle of *respect des fonds*, which demonstrated the imminent need to formulate concepts and principles that could theorize and rule existing practices in public and private institutions.

In the context of archival theory and practice, this principle was to support the processes of organization and representation of archival knowledge, acting as a guide
for the implementation of archival functions.

According to Duchein (1983, 64), “the simplest definition of respect des fonds means to group, without mixing them with others, the archives (documents of every kind) created by or coming from an administration, establishment, person, or corporate body”.

Heredia Herrera (2003, 5-6), talks about two principles: the principle of “Provenance”, which "means not mixing the records produced by different bodies" and, the principle of “Original Order”, which means not breaking with the natural order of records produced by different types resulting from the administrative procedures that determined it".

The principle of respect des fonds can therefore be understood at two levels: (a) records from the same producer cannot be mixed with those of another origin; b) records must be preserved in their original order, if it exists, and must reflect the order of the procedure that determined the production of the documents. This definition encompasses two levels of principle: provenance itself (the producer or receiver of the record) and original order (which defines that the organization of the producer must be maintained).

In addition, it was necessary to establish rules for actions and procedures adopted by professionals working with records. These rules were promulgated in manuals, which became the first scientific corpus in archival science and the basis for developing knowledge about archives and records. From these publications, archival science was to establish itself as a discipline, since the rules and methods postulated in them generated the first theoretical reflections on the area, enabling its addition to the disciplinary field.

In the mid-20th century, due to the need to rationalize the organization and deletion of records produced by public and private administrations, new concepts and functions emerged, such as the theory of (primary and secondary) records values, the difference between record and archive, records appraisal and, more recently, identification as a methodology for records management and handling of records accumulated in archives.

At present, in archival identification, studies of record origins are carried out based on the theoretical and methodological parameters of diplomacy.

Although diplomacy is older than archival science, its object, the documentary form, has potentially been confused with the archival record itself since the beginning of diplomatic studies. The relationship between the disciplines only became explicit in the late 1960s, gaining strength in the 1980s, when it became clear that information technologies and new forms of record production would affect the work of the contemporary archivist, leading to a paradigm shift in the discipline, according to Thomassen (1999).
Carucci (1987, 27) reinforces the idea of a common object between the disciplines, when she writes that "Diplomats studies the single record, or, if we prefer, the elementary archival unit, document[...] ... analyzing, above all, its formal aspects in order to define the legal nature of acts, whether in relation to their genesis or in relation to their effects".

Rogers (2015) understands diplomatic analysis as a process of abstraction and systematization, where the diplomatist deconstructs a document to identify and locate elements that reveal its provenance, relationships, reliability and authenticity.

In a new reality of information production, organization, and use, it is important to establish the reasons behind records creation, what the relationships are between them and their creator and what intentions lie behind the action of recording the information. To do so, analysis of the record, based on diplomatics and its method, has become one of the most secure ways for the archivist to understand it. Being able to establish who produced it, why and for what purpose, through the study of the documentary form is the great contribution of diplomatics to current archival studies.

Archivists have begun to understand the diplomatic method as a new tool to assist with record management in administrative processes. Diplomatics, which previously dealt exclusively with the documentary form, has now, as Bellotto (2004) points out, "broadened in the direction of documentary genesis and its contextualization in the attributions, competencies, functions, and activities of accumulating produced entities."

This new "use" of diplomatics by archival science has given rise to what is called "contemporary diplomatics", or "archival diplomatics", or "documentary typology".

Diplomats and identification: the production of archival knowledge

Bellotto (2004) made the distinction between the object of classical and contemporary diplomatics, called “documentary typology”, and their method of analysis, which can be complementary to archival identification.

The object of documentary typology is the elementary archival unit analyzed as a documentary genre and defined as the “configuration that assumes a record according to the disposition and nature of the information contained therein”. The typological identification has as its object the record type, reflecting the “organic logic of documentary sets”, which by definition is the “configuration that assumes a genre, according to the activity that generated it” (Bellotto 2004).

Rodrigues (2016), when studying the relationship between diplomatics and archival Science and its relevance to the studies of the origins of records emphasizes that "the need to identify records in its production context for planning its creation / production and treatment of its accumulation in the archives led the area to think over the identification as an archival process and the discussions about the place it occupies in the context of archival methodologies" (Rodrigues 2016).
Identification is a research phase, an intellectual task, the object of study of which is the produced entity and its records, whose recognition process is based on the parameters of diplomatics, in its classical and contemporary perspective - documentary typology - producing knowledge for the planning of activities that support archival functions (creation / production, classification, appraisal and description) (Rodrigues 2016).

The methodological basis of identification is the direct application of the principles of provenance and original order, it is its best tool (...). The characterization of the attributes of the creator, its functions and activities and the way in which these established links are translated in the parts of the archival records integrate the knowledge generated in this phase and make up the theoretical basis of the archival identification (Rodrigues 2002).

In this way, we predicate that the archivist, when conducting an identification study (of the creator and its records), based on the diplomatic and documentary typology, produces archival knowledge.

Tognoli et al. (2013) defined archival knowledge as all the knowledge produced about a particular person or entity and grouped into *fonds* (a unit consisting of the set of records accumulated by an entity).

From the examination of the parts, the archivist manages to arrive at the examination of the whole, of the context, of the process, in order to determine the relations existing between one and the other. The results achieved through the identification of the record, through diplomatics, will support the process of representing archival knowledge, in the context of classification and archival description.

**Diplomatics and description: contributions to the organization and representation of archival knowledge**

There are three functions related to organization and representation in archival science: classification/arrangement, appraisal and description, all functioning as product-generating processes and management instruments to control, access and accumulation.

Concerning appraisal, when the archivist needs to set deadlines for the elimination or preservation of records, the product will be a representative instrument of records values and destination.

In archival science, classification refers to the establishment of a logic for organizing the components of the fond, which is represented in the classification plan, a device that enables better apprehension of the institution’s operating context and facilitates the retrieval of information contained in documents.

The description aims to represent the content of archives and their records, recording elements of the representation of production contexts and content of the documentary series, whose instruments allow access to this information.
Among these two last functions, there are processes that aim to produce instruments both to organize and represent the content and context of records, such as a classification plan; a representative result of the condensation of institutional information; and internationally standardized descriptive representations for archival description or help finders (guides, inventories or catalogues).

Regarding classification, the application of diplomatics, through diplomatic analysis and typological record identification, has allowed archivists to establish creators and functions, which are necessary to prepare a classification plan.

In relation to the description, the specific theme of this paper, the methodological contribution of diplomatics becomes even more clear, especially after the moment of documentary analysis (when a synthesis is made of the elements extracted from the record to communicate its information).

Etymologically the word "describe" (derived from the preposition de and from the verb scribere) means "to write about". Therefore, the term "archival description" literally means writing about archival material, encompassing the ideas of identification, representation, and organization.

We will use these three ideas as the basis from which to highlight the function of description and its relation to diplomatics.

The first point to consider is the support given by diplomatics to what is called "documentary analysis", which, in the scope of archival science, means to initially extract the necessary elements for the identification of records. Here, the external and internal characters of the documents (recognized by the diplomatic method) are analyzed so that the archivist can identify their content which includes: genre, activity and provenance.

These elements can be extracted by applying the diplomatic method, where the provenance is identified in the initial protocol and/or in the escatocolo; the activity is identified in the dispositio, and the dates in the escatocolo or initial protocol. The genre is also identified through analysis as an intermediary element, that is, the record form as a suitable editorial structure for the insertion of specific content.

The second moment of "documentary analysis", the application of the diplomatic analysis to the record, consists in elaborating a summary of the record. According to Bellotto (2004), with the diplomatic method the archivist will perform a closer documentary reading, to get a necessary understanding through an effective typological analysis, involving such aspects as appraisal, classification and description.

This second stage consists of a representation, that is, from the extraction of the elements carried out in the identification stage, products that represent the record (its content and context) will be elaborated. These products are the help-finders or reference instruments that will explain records in terms of their location, as well as providing the
Organization constitutes the third stage of the descriptive process and concerns the disposition of the information previously identified and represented. At this point, the role of diplomatics is equally important, since its purpose is essentially to study the form of the document, its internal and external elements. The discipline will contribute to the structuring of help-finders, so that relevant information is available in its respective instruments; that, the form of the guide, inventory and catalog corresponds to the content of what it is intended to represent.

The theoretical-methodological support of diplomatics is clearly present in the descriptive process, since documentary analysis, subsidized by diplomatic and typological analysis, is prior and necessary to the archival description. Diplomatic analysis is investigative per se, since the analytic process – considered a process of abstraction – deconstructs the document to analyze its parts. This will reveal the most important elements, such as provenance and documentary type.

In the descriptive process, the archivist uses documentary deconstruction, appropriating the necessary elements in order to be able to represent the content and context of the information, systematizing it in the help finder.

Thus, it can be said that documentary analysis carried out on the basis of diplomatics will provide the elements necessary for the archivist to identify, represent and organize the content of the records and their set, the documentary series that integrates the groups and *fonds*, in order to guarantee the retrieval of information contained in these documentary sets and in their contexts, and so enable their exchange, diffusion and use.

**Conclusion**

The theoretical-methodological contribution of diplomatics and its method to archival functions within organization characterizes the discipline as formative to contemporary archivists who, in order to respond to the challenges posed by the increasing amount of records and information produced in different environments, have seen their role as guardians of historical documents being extended to record management.

For the purpose of description, diplomatics may subsidize documentary analysis, since its method is considered "an analytical-comparative method that makes possible the study of the relations of the documents with their creation context and with its producer, through the partition and the study of the form of the document, in order to understand it as a testimony of a fact" (Tognoli 2014, 104).

Besides the knowledge generated by record deconstruction, through diplomatic analysis, the archivist can now deal with the knowledge generated from the identification of information producers and their relationships with the procedure to represent the context and content of records. As stated by Rodrigues (2006),
identification as an archival methodology allows the normalization of the archival functions that support the procedures of records management, contributing to the organization and representation of archival knowledge.

References


Abstract
Instructors and students of World History worry about coherence: How to connect the study of diverse times and places so that students appreciate why these are studied in one course. This paper argues that the coherence problem in World History is for the most part a Knowledge Organization problem. It discusses how a set of KO strategies can be applied to the study of World History. Most obviously careful classification of the “themes” addressed in such a course allows students to better appreciate how themes interact through history. Visualization techniques can be particularly important in this task. We can also classify key agents and the key challenges they face. And we can draw upon KO understandings in addressing issues of multidimensionality.

Introduction and motivation
World history courses have become increasingly common in recent decades in many parts of the world. They reflect the twin desires to appreciate the connections between different regions of the world through history – how technologies and religions and stories and much more have crossed societal boundaries – and to also compare developments in different societies. Yet the very scope of world history has raised a challenge of coherence: Instructors and students may appreciate discussing Babylonia, Aztecs, and Polynesians in one course, but struggle to draw connections across time and place (Mitchell et al. 2016).

The challenge of world history is a Knowledge Organization challenge. How can knowledge about history – the names and dates, as well as our attempts to understand why particular events or processes occurred – be organized in such a way that it is easy(ier) to draw connections across different times and places? Moreover it is a particular type of Knowledge Organization challenge for our goal is to sustain student interest over a period of months: We must thus apply KO principles in a manner that is neither tedious for the student nor difficult to master.

The author is a scholar of both KO and history. A couple of years ago he became aware of the concerns within the world history community about the lack of coherence in world history textbooks and courses. He has since written a book and some articles in world history which apply KO principles to address this challenge. This paper represents the author’s first attempt to communicate these ideas to the KO community.

Themes, classification, and visualization
World history decades ago tended to stress cross-societal connections. In recent decades, though, a thematic approach has become common in which certain “themes” – culture, technology, environment, and so on – are traced through time, with
similarities and differences across regions being given special attention. These themes can each be seen to represent major subjects of scholarly interest, mostly within the social sciences and humanities. As such they tend to be prominent within KOSs: They are often main classes or important subclasses. KO can then play a critical role in clarifying terminology: what sort of subclasses does “culture” or “technology” embrace? A student who reads a 500-page textbook that follows such themes through history without ever defining them (an all-too-common practice unfortunately) will face unnecessarily great challenges in identifying coherence.

The simple recognition that themes interact through history – that culture shapes politics which facilitates certain economic behaviors which then affect the environment, and so on – opens the door to an even greater KO contribution. If we use KO to carefully identify both themes and subthemes then it becomes much easier to trace how themes and subthemes influence each other through history. If we define “institutions” so vaguely that the term subsumes most of what we call “culture” then it will be impossible to trace in an intelligible manner the ways that cultural attitudes and the formal rules of a society (that is, institutions) have influenced each other through history. And this, I would suggest, is one of the key lessons that a world history course can impart: Each “theme” may be studied primarily by one discipline today, but we cannot understand how “economy” or “politics” or “culture” or “technology” changed historically by studying these in isolation. But of course if our understanding of each theme is vague then we will not fully appreciate interactions across themes.

The Basic Concepts Classification (Szostak 2013) was grounded in a three-page classification of the key phenomena studied by human scientists formulated in Szostak (2003). That shorter classification can be used to clarify the main subclasses (components) of “institutions” or “culture” in a world history text. [Some students of History may then become interested in the more precise classification that BCC provides and thus in broader issues of Knowledge Organization.] The main classes and key subclasses of any classification system organized along logical principles might serve a similar purpose. Since a World History text cannot address every subclass for every historical episode addressed, such a table can also guide students to reflect on what may be missing from the text: This might help them formulate topics for writing assignments.

Note that the earlier (and still important) world history focus on cross-societal interactions is also enhanced by a superior understanding of thematic interactions. It is ideas and objects that are transmitted across societies: It is only through understanding thematic interactions that we can appreciate how these ideas or objects exerted effects on economy or polity or other themes. And our understanding of the importance of thematic interactions leads us to expect that changes in one theme will encourage changes in others.

One important topic within Knowledge Organization involves visualization
techniques: We appreciate, that is, not only that we need to organize knowledge properly but that we then also need to aid users visually in navigating our organization (e.g. Slavic et al. 2013). An emphasis on interactions among well-defined themes and sub-themes lends itself to a simple but powerful visualization technique: flowcharts that can identify the thematic interactions involved in any historical event or process. Any one such flowchart aids the student in making sense of that event or process: It gives them a structure on which to “hang” details (students need to be warned that the diagram is a complement to the surrounding text, not a substitute for it). Dozens of such flowcharts communicate powerfully the message that thematic interactions are of critical importance in history. And a series of flowcharts can communicate a cumulative story: the development of agriculture sets the stage for state formation in one flowchart while state formation has effects on international trade and other subthemes in later flowcharts. There is perhaps no better way to communicate to students that history is (in large part) a cumulative set of thematic interactions.

Figure 1 discusses the key causes and effects of the development of agriculture. Note that each rectangle in the flowchart identifies both a main theme and a subsidiary phenomenon. The diagram thus communicates powerfully how themes interact while nevertheless clarifying the nature of particular thematic interactions.

Figure 1: Influences on and effects of Agriculture
Note: The arrows on the left side capture how population pressure and/or climate change encouraged a gradual process of developing the technology for agriculture. This process was conditioned by the availability of appropriate plants and animals (and rainfall etc.). The arrows on the right capture how agriculture then encouraged political consolidation and stratification, war, further population increases (but reduced health), and together these generated a change in gender relations. Soil exhaustion is possible. Other effects could be added. Each box references both a general theme and a particular phenomenon within that theme.

While one flowchart can powerfully communicate the importance of thematic interactions in a particular historical “episode” such as the development of agriculture, a series of flowcharts can communicate the importance of thematic interactions through time. Many of the phenomena captured on the right hand side of Figure 1 – urbanization, male power, war, states – will play important causal roles in later flowcharts in a world history text. Students can then appreciate one of the key lessons that a world history text should seek to impart: that history is a cumulative process of thematic interaction (see Reilly 2013).

**Challenges facing historical agents**

It is often noted that students tend to personalize history: They can be taught about how technological innovation or climate change exerted particular effects but on an exam will struggle to attribute the results to individual human agency. One of the tasks of world history is to communicate an understanding of how human agents interact with society-level forces. If we squeeze individuals out of historical accounts – which is easy to do when trying to communicate the history of the world in a few hundred pages – then students will insist on putting them back in. Though we may struggle to give particular individuals a lot of space in a world history course, we can lend further coherence to a world history course by discussing the common challenges faced by various types of individual through history. Farmers have to worry about theft, rulers have to confront bureaucratic corruption and disobedience, and merchants need to worry constantly about both theft and deception. Much of world history can be understood as either attempts to address such challenges or failures to do so. We can then easily compare how people in different societies addressed these challenges. The fact that some rulers spied on their bureaucrats, and some of these killed the families of bureaucrats that (seemed to have) robbed them are not just historical curiosities but reactions to a common challenge facing rulers: We can then explore why different rulers made different choices and what the effects of these choices were. And students can readily compare the actions of rulers – or farmers, merchants, parents, and a host of other types of people – in quite different times and places. And we can also trace how people developed solutions to these challenges over time: Merchants developed networks that threatened to expel members who robbed or deceived other members,
and such networks facilitated trade expansion.

Note that we need first to identify key “types of individual.” This involves thinking about the various roles that humans play in society: Occupations loom large here but other roles such as parents or voters deserve attention. And then we need to classify key challenges: This requires reflection both on the goals of particular actors and how they interact with others.

We also communicate another important lesson of World History: that similar challenges have been faced across many times and places and often been addressed in a similar manner.

One key advantage of identifying common challenges faced by different types of human agent is that we then have a set of common questions to ask any time that type of actor appears in world history. The same can be said of evolutionary analysis: It guides us to ask about sources of mutation (in culture, art, technology, science, or institutions), the selection environment for these, and transmission mechanisms across generations. It thus also facilitates comparisons across time and place. And as with common challenges it helps us also to understand change through time: We can see how particular mutations built on what existed before, and study how ideas are transmitted across space and time.

**Coping with multidimensionality**

One of the challenges of world history is its multi-dimensionality: We want to simultaneously understand changes through time, connections across societies, and interactions among themes. As the KO community has long appreciated, the same “fact” may have a place within different organizing schema. The world historian should thus be conscious that in emphasizing any one of these dimensions we will inevitably divert attention from others. World history texts tend to organize chapters chronologically. This can potentially allow the author to explore key thematic interactions or societal connections at a particular point in time. But such an approach limits the author’s capacity to make comparisons across a broader set of times and perhaps places (for many world history texts discuss different regions in different chapters) than a particular chapter encompasses, and also limits the author’s ability to describe historical processes that transcend a particular period of time (though we saw above that flowcharts may allow students to better trace a complex historical process across chapters). One possible approach here is to use in-text boxes to purposely transcend the boundaries of a particular chapter: to purposely compare a particular thematic interaction (say food shortage and political insurrection) across quite different times and places, or to trace a particular historical development across several time periods. We can, for example, provide a quick history of rubber in one place – perhaps when the Spanish send rubber balls back to Europe in the 16th century to amaze the aristocracy – which can show how both various themes (war, trade, culture, technology)
and different types of people based in different regions (Aztec ballplayers, Spanish conquerors, merchants, plantation owners and workers in Africa and Asia, scientists and innovators in many countries) interacted over a period of centuries as rubber moved from use in games to becoming a key product of the industrial/automotive age. The history of rubber thus serves as a microcosm of world history itself.

Skill acquisition

We have spoken a couple of times above about how KO techniques help us to communicate key lessons of World History. The same strategies aid skill acquisition also. Most obviously, students should learn how to better organize a large mass of information. This is a critical skill in our present age of “information overload.” Yet we also noted above that students can better appreciate thematic interactions if themes are first carefully defined through classification. The act of organizing thus facilitates critical analysis itself. Quite simply, students must first understand the nature of the themes they study before they can hope to comprehend what happened in history. The better they understand themes, the better they can analyse arguments involving those themes. Last but not least, our use of visualization techniques should enhance what is often termed visual literacy.

Concluding remarks

Knowledge Organization has a role to play far beyond our libraries and online databases. This paper has argued that Knowledge Organization can revolutionize the study of history. It likely has a similar role to play in many other disciplines (perhaps especially

References
Approaches to the concepts of exhaustivity and specificity in ISKO International meeting proceedings: 2000-2017

Abstract

We study how the concepts of exhaustivity and specificity are addressed in the publications of the ISKO international meeting proceedings for the period 2000-2017. In particular, we analyze the aspects related to thematic proximity and the methodological approaches in these publications. For the selection and analysis of the corpus we used an ad-hoc combination of techniques and methodological procedures, including content analysis and analysis of keywords in context. The results show that studies on exhaustivity and specificity are scarce and not very central in ISKO meeting proceedings, while of the most of the publications follow empirical approaches.

Introduction

The origins of subject indexing, as a technique for document analysis, date back to ancient times and civilizations such as Mesopotamia (Witty 1973). Subject catalogs, whose origins are also rooted in these ancient civilizations, were established and systematized with the work of Charles Ammi Cutter, whose “Rules for a printed dictionary catalog” (1904) provided guidelines for this practice (Witty 1973; Pettee 1945; Borko and Bernier 1978). In this vein, the ultimate goal of the operation of indexing in a catalog, since it is based on the analysis and representation of contents using indexing terms (Rowley 1982; Chaumier 1982; ISO 5963:1985), would be information retrieval.

In the operation of indexing there are two key principles that affect the retrieval of documents: exhaustivity and specificity. The concept of exhaustivity is related to the number of subjects (factors) that are translated to concrete representative terms for a document. According to Jones (2004), the exhaustivity of a document description is the coverage of its various topics given by the terms assigned to it. For Khosh-Khui (1986), exhaustivity refers to the extent to which a document is analyzed to completely identify its contents. Anderson (2002) sees exhaustivity as the number of single concept terms that will be used to describe the topics, content or meaning of a documentary unit. For Olson and Given (2003), exhaustivity would be the breadth of representation, the number of factors indexed and concerned with the different aspects included, which leads to comprehensiveness. In this sense, exhaustivity is related to the level of indexable matter and how much of a given topic must be covered. According to Ogilvie and Lalmas (2006), exhaustivity measures how exhaustively an element discusses the topic.
Specificity, on the other hand, is the level of detail in which a particular concept is represented (ISO 5963: 1985). According to Jones (2004), the specificity of an individual term would be the level of detail at which a given concept is represented. For Khosh-Khui (1986), specificity is the extent to which a system allows precision in explicitly stating the subject contents of a document. According to Anderson (2002), specificity refers to the tightness of fit between the meaning of a term and the topic of a discussion or illustration in a text, which tends to offset or modify the impact of exhaustivity. According to Olson and Given (2003), specificity is the relative detail within the vocabulary, the number of hierarchical levels defined, which increases with each level as the hierarchy becomes deeper. In this way, the level of specificity should vary according to the subject itself and the users’ needs. Mai (2004) also talks about exclusivity, stating that classes on the same level should be distinct, such that documents placed in one class could not also be placed in another class. According to Ogilvie e Lalmas (2006), specificity measures how focused an element is on the topic.

According to some authors (e.g., Foskett 1977; Chaumier 1982; Langridge 1989), these two principles would be political/administrative decisions as they are based on the indexing policies of their respective information services. The different degrees of exhaustivity and specificity in the indexing of documents affect the relevance of retrieved documents, as a greater number of appropriate index terms increases the density of documents judged as relevant (Wolfram and Zhang 2002; Kim 2006). On the other hand, the concept of relevance depends on external factors, such as the user’s background knowledge (Abdulahhad et al. 2013). According to Pehcevski and Larsen (2007) relevance is the extent to which some information is pertinent, connected, or applicable to the matter at hand, representing a key concept in the fields of documentation, information science, and information retrieval.

Given and Olson (2003), in a more critical way, also cite these principles as important strategies for effective information organization and retrieval in research, also relating them to the concepts of recall and precision (an aspect introduced by Richter 1984). For these authors, recall would be the amount of relevant information that is retrieved, while a maximum level of recall would mean “retrieving every last instance of a theme or variable. [...] If exhaustivity is high more codes are used, which will allow more data to be retrieved and analysed” (Olson and Given 2003, 131).

Specificity, on the other hand, would be related to precision. Olson and Given (2003, 130-131) point out that “precision is enhanced by high specificity, [as] narrower categories will produce fewer data in each category.” Thus, “if precision is high, then all the information retrieved is relevant and little or no irrelevant information is retrieved.” Thus, greater exhaustivity in indexing entails a high level of recall and a low level of precision; while greater specificity entails a high level of precision and a low level of recall. This means that recall and precision tend to be inversely proportional, which will have "an impact on the construction of data categories and codes for
analysis” (Olson and Given 2003, 131). According to Khosh-Khui (1986), exhaustivity controls recall potentialities, and specificity controls precision capabilities. Thus, it is said that exhaustivity “is recall-oriented,” while specificity is “precision-oriented” (Abdulahhad et al. 2013).

Abdulahhad, Chevallet, and Berrut (2013) state that exhaustivity and specificity are “still theoretical notions without a clear idea of how to be implemented” and, in a logical framework of practical application, conclude that “the explicit integration of Exhaustivity and Specificity into IR models will improve the retrieval performance.”

In the words of Anderson (2002), “the detail of indexing is a very important policy consideration for any index, with strong implications for search precision and recall.” Thus, the principles of exhaustivity and specificity are essential in the process of indexing as they are closely related to information retrieval.

Evangelista, Simões, and Guimarães (2016) conducted a study on these principles as ethical values in knowledge organization. The authors identified how these principles affect information retrieval in relation to the diversity and expressiveness of indexing terms. Regardless of the nature of these two principles, it is well-accepted that they influence the entire dynamics of the indexing process, especially in relation to the analysis, representation, and retrieval of information. As a consequence, it is also well-accepted that these two principles should be considered in an “objective” way in order to increase precision and recall.

Considering that exhaustivity and specificity are two of the main concepts of subject indexing, this study aims to analyze how these two concepts are addressed in the publications of the ISKO international meeting proceedings for the period 2000-2017. In particular, we aim to analyze the aspects related to their thematic proximity and methodological approach. Our objectives are: (i) to identify and systematize the elements that characterize the conceptual construction of the principles of exhaustivity and specificity; (ii) to identify the works on the topic in the ISKO international meetings proceedings (2010-2017); (iii) to identify and describe the thematic proximity and methodological approaches to these two concepts in the papers.

**Methodology**

For the selection and analysis of the corpus we used an ad-hoc combination of techniques and methodological procedures, including content analysis and analysis of keywords in context (Bardin 2011; Bernard and Ryan 2010; Coutinho 2013).

Regarding the first objective, we conducted a literature review of the topic. In relation to the second and third objectives, we searched the terms “exhaustivity,” “specificity,” and other related concepts such as relevance, precision, recall, and consistency in the full texts of the ISKO international meeting proceeding (2000-2017) and analyzed the meaningful parts of the papers (mainly the title, abstract and introduction).
For the analysis of the methodological approach, we adopted two nominal variables (see Table 1), based on Bernard and Ryan (2010, 151).

Table 1: Variables used in the analysis category “methodological approach”

<table>
<thead>
<tr>
<th>Methodological approach of the paper (description)</th>
<th>Variables of the approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>The authors analyze the theoretical-methodological foundations and discuss the relevance of exhaustivity and specificity in indexing (especially for the case of information retrieval) exclusively based on the analysis and information of texts.</td>
<td>Epistemological analysis</td>
</tr>
<tr>
<td>The authors present results on the relevance of exhaustivity and specificity in indexing (especially for the case of information retrieval) based on experience.</td>
<td>Empirical study</td>
</tr>
</tbody>
</table>

In order to study the thematic proximity of the papers to the concepts of exhaustiveness and specificity, we worked with four ordinal variables that were expressed in a negative scale of intensity (see Table 2), based on Bardin (2011, 84).

Table 2: Variables used in the analysis category “thematic proximity”

<table>
<thead>
<tr>
<th>Thematic proximity of the paper to the concepts of exhaustivity and specificity</th>
<th>Degree of proximity</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>The concept is a core part of the study.</td>
<td>Central</td>
<td>0</td>
</tr>
<tr>
<td>The concept is addressed because of its intrinsic relationship with the object of study.</td>
<td>Inherent</td>
<td>-0.5</td>
</tr>
<tr>
<td>The concept is addressed because of a secondary relation with the object of study.</td>
<td>Peripheral</td>
<td>-1.5</td>
</tr>
<tr>
<td>The concept is not addressed, but we infer a thematic connection.</td>
<td>Inferred</td>
<td>-3</td>
</tr>
</tbody>
</table>

Results

Out of the 578 papers that were published in the period 2000-2017, we selected and retrieved 29 papers (5% of the total) that are relevant for analysis. Figure 1 shows the frequency of these publications per year, noting that most papers were published in 2000 and 2002. Since then, there has been a significant decrease in publications on the topic until 2012 (the year with the lowest frequency), and again a mild increase in 2014 and 2016.
Out of the 29 papers, only seven papers (24%) worked with the concepts of exhaustivity and specificity (see Table 3). The other 22 papers (76%) address related concepts such as relevance, precision, recall, and consistency.

Table 3: Papers working with the concepts of exhaustivity and specificity

<table>
<thead>
<tr>
<th>Author/Year</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frâncu, V. (2000)</td>
<td>Harmonizing a universal classification system with an interdisciplinary multilingual thesaurus: advantages e limitations</td>
</tr>
<tr>
<td>Broughton, V. (2002)</td>
<td>Facet analytical theory as a basis for knowledge organization tool in a subject portal</td>
</tr>
</tbody>
</table>

Related to the methodological approaches, the majority of papers (26, 89%) are empirical studies, while only 3 papers (11%) followed an epistemological approach.
Most of these empirical studies focused on the use of classification systems or information retrieval systems such as OPACs, etc. and also mentioned the concepts of recall, relevance, and precision.

As for thematic proximity, only one paper (3%) – Chen 2008 – was central in dealing with the concepts of exhaustivity and specificity. As for the other papers, the degree of proximity was inferred in 22 papers (76%), and peripheral in 6 papers (21%). We did not identify any paper with an inherent degree of proximity.

Conclusion

Based on the results, we conclude that studies on exhaustivity and specificity are scarce and not very central in ISKO meeting proceedings. Most articles do not deal with indexing as a whole and very few specifically address these two concepts. Given that specificity and exhaustivity are present not only in the process of indexing but also in a wider range of areas related to the representation and retrieval of information, we believe that more research is needed. The clarification and study of these two concepts is important as they are also related to other concepts that contribute to semantic ambiguity, such as consistency, precision, recall, and relevance.

The prevalence of empirical approaches in the methodologies, such as case studies and applications of theoretical models, might be due to the fact that the studies of concepts such as precision, recall, and relevance, generally associated with the concepts exhaustivity and specificity in information retrieval, mainly follow empirical-methodological approaches rather than theoretical-methodological and epistemological approaches. However, the small number of studies on exhaustivity and specificity in the proceedings, and the inferred or peripheral centrality of most of them, might indicate that the relationship between knowledge organization and information retrieval is an under-researched area that deserves more attention from the ISKO community.

Acknowledgment

Isadora Victorino Evangelista thanks Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) 2017/02327-8 for its support.

References


Gregory H. Leazer, Robert Montoya, Jonathan Furner

Articulating a cultural research program for Knowledge Organization Systems

Abstract
Classification and knowledge organization (KO) are increasingly ubiquitous in everyday life. We briefly examine two theories of KO development, the theory of scientific classification, and the theory of technological progress, and assess the subjective cultural aspects of classification. We find that the cultural dimensions significantly shape and permeate classification and KO activity, and we call for basic research investigation into the cultural dimensions of KO. Such investigations fall into three research fronts: rational, empirical and ethical. Finally, we call for a critical program that supports users efforts to understand the operations of KO systems.

"Knowledge organization (KO) system" and related terms (such as "classification," "taxonomy," "ontology" and "vocabulary") pick out members of a family of concepts whose referents share characteristics of several kinds. The referents vary from Internet search engines like Google, proposed systems like the Semantic Web, and international systems for bibliographic control, to systems like Facebook, which serve as platforms for personal interaction and as sources of shared reading, and which have similar functions. Such systems are increasingly characterized as "knowledge infrastructure" that mediates access to (by various formulations) information, knowledge, or documents. They comprise and are constituted by various standards, operating in multiple kinds of inter-institutional practice, from local practice to highly integrated global arrangements.

KO and classificatory activities take many forms: rankings, orderings, set definitions, descriptions, metadata, and the web of associations amongst terms, sets, and objects. Many of the conclusions from the field of critical algorithmic studies apply to our understanding of classificatory activities: they are ubiquitous, and closed; they incorporate and reiterate forms of social bias; and they can produce real social consequences for the people that interact with or are subject to them. Many of these items are discussed, for example, in Bowker and Star (2000), Pasquale (2015), and O'Neil (2017). KO is well-defined as "the field of scholarship concerned with the design, study, and critique of the processes of organizing and representing documents that societies see as worthy of preserving" (Tennis 2008, p. 103), but, because of the ubiquity of new media, we must view classificatory and KO activities as extending beyond information retrieval per se. KO systems such as Amazon's and Netflix's recommendation systems increasingly profile and classify users by a variety of markers, and adjust document retrieval sets accordingly. Tailoring online advertising to users can be conceptualized as an information retrieval problem, although it is not typically
considered to be a mainstream KO problem by the KO community. Typically considered to be even farther afield are the problems addressed by automated bank loan approval systems, college admissions systems, and parole approval systems based on estimations of recidivism. All are classificatory systems, operate similarly to automated KO systems, and provoke many of the same concerns with their use.

The last century of development of classifications, and of KO systems more generally, has proceeded with two major assumptions:

- KO systems should be modeled after scientific knowledge, using scientific processes that are attentive to expert consensus and evidentiary practices, to develop their conceptual content and structural arrangement. We call this the theory of scientific classification.
- Advances in KO system research are due to a functional experimental program that has relied upon technical advances. These techniques include the development and use of controlled languages, computer-based search and retrieval, and more recently increasingly sophisticated statistical approaches. We call this the theory of technological progress.

The field of KO system development is often conceptualized as a technological endeavor that has shifted, from a field where documents were represented by metadata gathered through manual processes and browsed in relatively simple manual systems, to a field where documents are represented by automatically generated metadata and searched in computational systems. Key to this transformation has been the implementation of the computational algorithm and increasingly sophisticated statistical approaches.

This technological change has occurred in a period of rapid social change. Globalization has further imperiled indigenous cultures and languages; increased social stratification; and digitization has not only multiplied documentary genres and formats but also enabled concentration of media ownership. These trends, along with increasingly onerous copyright and intellectual property regimes, have restricted fair and open access to knowledge resources. KO, once the exclusive purview of universities, libraries, archives, museums, and learned societies, with some commercial support, now is significantly shaped by major commercial players like Google, Thomson Reuters, and Reed Elsevier.

However, in a period of rapid technological and social change, key assumptions regarding the nature of the human mind, of sociocultural relations, of classificatory structures, and of linguistic phenomena remain relatively under-investigated. A lack of understanding of the social and cultural practices that underpin human interactions with information and with KO systems has led to undesired consequences in the design, development, and use of KO systems. KO systems, by failing to account for various kinds of social difference and variation, incorporate social bias, and reinforce it.

The theory of scientific classification had H. E. Bliss as one of its most significant
early advocates, particularly with his book of 1929, *Organization of Knowledge and the System of the Sciences*. However, later theorists have emphasized the lack of consensus in scientific exercise. Kuhn's *The Structure of Scientific Revolutions* (1970) is an example of the myth of consensus; Hull (1988) and Gould (1981) demonstrate the difficulties of, and mismatches between, human categorical modeling of the natural world and scientific taxonomy. The scientific approach is idealized, and cannot avoid the kinds of distinctions that are rooted in cultural systems.

There is a tradition of examining cultural bias present in classification. In the past, scholars like Hope Olson (2002) and Sanford Berman (1971) concentrated on problems of nomenclature, particularly around the names of kinds of people and cultural practices, with special attention to gender and sexuality. Work by Melissa Adler (2017) is a more recent addition to this topic. Elaine Svenonius (2004) examines theories of language and their connection to the validity of knowledge representations as a more basic examination of how those theories limit or enable our conceptualization of classifications.

Automated classification systems have traditionally used term-frequency measures to determine the topicality of a document, on the assumption that terms that are relatively low-frequency across a collection, but relatively high-frequency within a given document, are good descriptors of that document. Clearly there are limits to this theory, similar to those "postulates of impotence" described by Swanson (1988), who states "word-occurrence statistics can neither represent meaning nor substitute for it. Such data, however, can be used, with occasional success, to signal or point out potentially fruitful areas of text…" (p. 95). Another potential direction of research would be to find out in which ways term frequencies fail, such as in the non-figurative or poetic language associated with lyric and poetry, and in fictional works. In such ways, we in the KO community might develop a greater sensitivity toward the limits of language in classificatory activities, even coming to an understanding that language itself is a cultural and social system; we might start to examine the ways in which language and our models of language limit, influence, and bias our KO systems, and the manner in which language does not present itself as a neutral field of scientific application.

KO systems are clearly incorporating a wider variety of data in their modeling of users and information. We are all familiar by now with Google's network-based centrality measures as a particular kind of citation practice, and the cultural (including science-as-culture) dimensions of citation behaviors as documented by Price (1963), Kaplan (1965), and Merton (1968). Other behaviors, such as shopping and viewing histories, location and search history are being incorporated in advanced retrieval systems. While these behaviors are strictly speaking non-linguistic, they are cultural signifiers, and hence semiotic.

The assumptions enumerated above can even be usefully examined within the field
of biological classification, where rational and empirical approaches are generally assumed to be part-and-parcel of taxonomic work. Scientific classification has been the model *par excellence* for all of classification. One major operation has been to determine the membership criteria by which a specimen could be assigned to a class or taxon. Far from being an accurate reflection of a naïve natural reality, scientific classifications again incorporate scientific and cultural expectations, which can act as a set of biases on those classifications. For centuries, biodiversity work has involved the production of taxa based on the possession of certain characters or organismal traits (Queiroz and Gauthier 1992, p. 452). However, identification of the attributes and features of an organism that are used to determine membership within a particular class is, by definition, a subjective assessment. This is a fact not only accepted in the biodiversity world, but also expected as part of the practice of building taxonomies in general—taxonomies are hypotheses, always-already situated, to be rearticulated and changed with the introduction of new scientific insights and methods. Scientists create classes; classes do not exist as natural kinds. Alongside these practices, biodiversity work has seen an increasing reliance on statistical and automated methodologies that have begun to obfuscate the individual judgment required in any and all taxonomic work.

In contemporary biodiversity practice, it is often the case that groups of organisms are computationally clustered based on molecular similarities. These so-called phylogenetic analyses rely on the examination of molecular traits that can be "measured and scored" (Felsenstein 1983, p. 247) and statistically evaluated to produce evolutionary trees of the most beautiful complexity. Genetic markers in the form of DNA barcodes (such as the mitochondrial C01 gene sequence) have been increasingly useful in constructing these schematics (Waterton, Ellis, and Wynne, 2013; Erickson and Driskell, 2012). The outputs of such phylogenetic examinations are drawings that depict the hypothesized relationship between various entities. One result of this approach is a "proliferation of taxonomic categories" (Queiroz and Gauthier 1992, p. 457). Alongside these approaches we see an overall decline in the use of morphology as the primary tool for assessment (Wheeler 2004, p. 571). The problem with this trend is that human judgment is often overlooked as still being essential to taxonomic practice. Individual intervention is still necessary to create classes and taxon groups. Even with the assistance of computational analysis and clustering, we often underestimate the cultural qualities of these classifications.

**Investigating the cultural elements of Knowledge Organization**

The age of the citations above indicates that investigations of KO as a kind of cultural practice have been in existence for decades. Nor has the presence of such research prevented technological development within the field. However, the presence of cultural elements within classification raises two kinds of broad concerns: first, does
cultural variability in classification pose a limit to the development of KO systems; and secondly, does the presence of cultural perspective within a classification scheme amount to bias or other kind of ethical malpractice?

We thus see investigations of classification and KO as reflecting a basic and applied research paradigm that is common in scholarship and in science. Cultural investigations – both in how classifications are made of cultural presumptions and also in how classifications reinforce cultural presumptions – form a basic research paradigm for classification because of how they may limit or reshape classificatory practice. Technological advances may continue, albeit we hope with a better critical understanding of the nature of classification, and with the development of more sensitive, more robust, and more ethical KO systems.

Within the basic research paradigm, we also envision investigations of a series of assumptions regarding KO. We have organized these assumptions as rational, empirical, and ethical research fronts.

**Rational assumptions**

1. Platonism about concepts as abstract objects. Occurrences (e.g., utterances and inscriptions) of terms, keywords, concept labels, subject headings, topic statements, classmarks, etc., are concrete particulars instantiating abstract universals known as concepts; and these concepts exist in the real world.
2. Pluralism about instantiation. It is possible for a single term, subject heading, etc., simultaneously to have multiple meanings (i.e., instantiate multiple concepts) for different people and in different domains, disciplines, communities, cultures, worldviews, etc., and for the same term to have different meanings at different times.
3. Structuralism about semantic relationships. It is possible, at least in principle, to devise conceptual structures whose internal relationships are sufficiently coherent to form the basis for controlled vocabularies that, whether they exhibit polyhierarchy (many-to-many concept–concept relationships) or not, avoid circularity.
4. Intersubjectivism regarding the boundaries of KO and KO system concepts. The concept knowledge organization system is itself a social construct.
5. Skepticism about objectivity in KO system design. All KO systems necessarily reflect (to a greater or lesser extent) the worldviews of their curators. Indeed, all KO systems reproduce (to a greater or lesser extent) the power relations at work in their creation.
6. Relativism about the instrumental value of KO systems. Such systems are culturally-specific artifacts. Notwithstanding their generic similarity, individual systems have different uses, functions, purposes, etc., depending on the contexts in which they are produced and consumed. Likewise, it is possible for a single
KOS simultaneously to have value of multiple different kinds to different people; and different KOSs may well be incommensurable. "Universal" KO systems can never be truly universal. Empirical evaluation of all methods, approaches, schemes, systems, etc., is always necessary; yet the choices to be made among evaluation criteria, warrants, authorities, etc., are always more-or-less arbitrary (or, alternatively, "culturally situated").

7. Like culture itself, KO systems are historically situated. The linguistic, social and conceptual spaces that are modeled in KO system are subject to change.

**Empirical assumptions**

1. The universalist principle. Harmonization, standardization, interoperability, linked data, etc., are good; disconnectedness, localism, silos, etc., are bad.
2. The techno-optimist principle. Large-scale classifying, facet-decomposing, harmonizing, etc., can be done automatically, to at least as high standards as those reached by manual means.
3. The pragmatic principle. In the creation of vocabularies, a combination of top-down and bottom-up methods is always better than either one or the other.
4. The anti-enumerative principle. Organizing concepts by facet is always best.

**Ethical assumptions**

1. KO system designers should strive for neutrality and work to minimize the impact of personal biases.
2. KO systems should be inclusive: they should meet the requirements, respect the identities, and promote the interests of all of their potential users, not just the "typical" user, nor an elite, nor the "majority".
3. KO systems should distribute terms, classes, etc., equitably: i.e., in ways that accurately and fairly reproduce the cumulative significance of the concepts represented for the whole user population.
4. Along with access, understanding, etc., transformation (particularly of power relations) and freedom from oppression are legitimate goals of KO.
5. Diversity is good; homogenization is bad.

Clearly, no single person could make all of these assumptions without facing several logical contradictions. We argue that it is just those contradictions that the KO community needs to address directly, and either eliminate or resolve, if it is to make substantive progress beyond the current state of the art.

**Developing a critical approach to the organization of knowledge**

Of increasing significance, then, is that KO itself must become the locus of critical reading and literacy. Users must not only develop an information literacy, defined as
"the ability to locate, evaluate, and use effectively the needed information" (American Library Association 1989), but develop a critical capacity for understanding how the various systems of knowledge organization permit, limit, or shape their means of access to information. Common conceptualizations of critical literacy serve as a point of departure for developing a notion of critical information literacy, especially as it pertains to KO. Critical literacy, as derived from Freire (1970), emphasizes the reader's ability to understand and analyze a text's ideological meanings and its role in various discursive formations, and that the reader and the text are both historically constructed. We seek a similar understanding not just of texts or of information, but of the methods of access to texts and information as well, and an understanding that the information seeker is an agent within a larger cultural milieu.

The consequence of enabling a more critical engagement with information systems entails exposing their functions, design assumptions, and biases to critical understanding. How to accomplish this task is also an open and relatively unexplored research question, but has received some attention, for example, in Drabinski (2013). Critical approaches will help us to establish more ethical approaches to the organization of knowledge, and to provide better articulations of their purpose and contributions to society.

References


Brisa Pozzi de Sousa, Cristina Dotta Ortega

Aspects regarding the notion of subject in the context of different theoretical trends: teaching approaches in Brazil

Abstract
Teaching in the area of knowledge organization is related to the construction of records in catalogues or databases, bringing together certain elements, such as the notion of subject, into a field designated as “subject information organization”. This study thus aims to observe the notion of subject, as an element of knowledge organization, in Brazilian teaching, according to three different theoretical trends: “subject cataloguing”, “indexing” and “documentary analysis” (“analyse documentaire”). With respect to these three trends, subject information organization was explored on some occasions during three graduate courses and in 13 topics, following meetings between directors and teachers of library science schools from Mercosul countries. We conclude that the notion of subject should be more deeply studied considering the marked differences between the three theoretical trends mentioned, and in order, among other things, to better understand and promote curriculum proposals, which should be based on theoretical and methodological keystones, in place of aspects of teaching focused on the functioning of widely known tools.

Introduction
The field of knowledge organization (KO) involves understanding the theoretical fundamentals related to the notion of subject in the context of subject information organization (in Portuguese, tratamento temático da informação – TTI). Esteban Navarro and García Marco (1995) say that KO aims to study the fundamentals and planning techniques, construction, management, use and evaluation of description systems, cataloguing, arrangement, classification, storage, communication and document retrieval. Regarding document content, in relation not only to structure but also to the subject – being the second the interest of this study –, information production is ensured through the documentary record, enabling the generation of new knowledge.

KO is considered as part of library and information science (in Portuguese, biblioteconomia e ciência da informação), and in Brazil it is usual, though not the consensus, that library science is taught at undergraduate level while information science is taught at graduate level. In this study, we considered the necessary dialogue and connection between these two academic levels, acknowledging that teaching of library science in undergraduate courses is bolstered by graduate activities, and indeed that the opposite also occurs. The first Brazilian master’s courses in the field emerged in the early 1970s and were designated as “library science”, nowadays referred to as “information science” – with the exception of the first course in the area to be developed in Brazil, at the Instituto Brasileiro de Bibliografia e Documentação (IBBD), now Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) (Mueller 1985).

In general, the name “library science”, in Brazil, is used to refer to the profession
and to undergraduate courses, as listed below.

One of the forums for debating areas of research in the field of KO in Brazil is the National Meeting of Research and Post-Graduation in Information Science (Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação – ENANCIB), set up by the National Association for Research and Post-Graduation in Information Science (Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação – ANCIB). According to Barreto (2009, 13), the ANCIB was founded in 1989, in Brasília, with the aim of: “promover o desenvolvimento da pesquisa e de estudos avançados da Ciência da Informação e Biblioteconomia no País”¹, as its founding statute indicates, approved by the General Assembly on June 23, 1989, at the X National Meeting of Postgraduate Courses in Library Science, Documentation and Information Science (X Encontro Nacional de Cursos de Pós-graduação em Biblioteconomia, Documentação e Ciência da Informação). Nowadays, the ANCIB website states that “sua finalidade é acompanhar e estimular as atividades de ensino de pós-graduação e de pesquisa em Ciência da Informação no Brasil”² (National Association for Research and Post-Graduation in Information Science – Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação –, online). Notice that the letter B remains in the acronyms for both the association and the meeting but is no longer present in the extended names.

Besides issues of naming, the aim of this study is to analyze the notion of subject, as an element of KO, in Brazilian teaching, according to three different theoretical trends: subject cataloguing, indexing and documentary analysis (analyse documentaire). Therefore, in view of these three trends, subject information organization (TTI) was explored on some occasions during three graduate courses and in 13 topics, following the meetings between directors and teachers of library science schools from Mercosul countries³.

**Construction of Knowledge Organization and the Brazilian context: three theoretical trends regarding the notion of subject**

At an international level, Ingetraut Dahlberg, in 1977, founded the Society for Classification, aimed at involving researchers interested in classification in both theoretical and practical study. It discussed development of formal and mathematical methods to recognize classes of objects according to their characteristics, using tools

---

¹ “promoting the development of research and advanced studies in library and information science in the country”.

² “its function is to accompany and stimulate the activities of graduate teaching and research in information science in Brazil”.

³ Curriculum studies in library science in Mercosul (in Spanish Mercosur) countries began following Program III of the Mercosul Plan for Regional Educational Development, which has involved joint cooperation of the Ministries of Education from Argentina, Brazil, Paraguay and Uruguay for some events.
like the thesaurus, for example (Dahlberg 1995). However, the studies of the Society for Classification were guided by the interests of one major council group – namely that of mathematics – and, due to this, Dahlberg left the Society to launch ISKO in 1989, remaining its president until 1996. Dahlberg’s conception led to KO being considered as an autonomous science (Dahlberg 2006).

Taking the international scene as an example, in Brazil, studies on KO were led by bibliographic classification, as in the specialization course in Information and Documentation, created in 1955, at the IBBD. However, according to Sales (2016), intensification of KO activities has occurred due to two main events: the creation of the first information science graduate course in the country, in 1972, at the IBBD, nowadays IBICT; and the launching of ANCIB, in 1989, in particular the work of research group number two (GT2), now called Organization and Representation of Knowledge. It was at ENANCIB, in 2007, in the context of the activities of GT2, that the chapter of ISKO-Brazil was established.

On a wider scale, the development of the first master’s courses in the 1970s, and the creation of ISKO International and of ANCIB, in 1989, led to more significant areas of discussion within KO.

In the context of these discussions about KO, Guimarães (2008, 2009) indicates that Anthony Charles Foskett was responsible for the rationale of the epistemological universe which, in Brazil, was designated “subject information organization” (tratamento temático da informação – TTI) in Brazil. According to Guimarães (2008, 78), this name became known from the 1970s onwards, especially because of Foskett’s book, “The Subject Approach to Information”, first published in 1969 and translated by A. A. Briquet de Lemos as A abordagem temática da informação (Foskett 1973). Guimarães (2008) mentions the influence of the expression in the teaching of library science in Brazil, when the designations “thematic representation” and “descriptive representation” substituted the names “classification” and “cataloguing”, within the framework of the Library Science Minimum Curriculum of 1982.

Guimarães (2008, 2009) points out that during the history of KO, there have been three distinct theoretical trends focusing on thematic aspects of KO, namely the North American “subject cataloguing”, British “indexing” and French “analyse documentaire” (in Portuguese referred to by the author as “análise documental”). Thus, “[...] hoje encontram importante espaço de interlocução no âmbito da

---

International Society for Knowledge Organization (ISKO)\(^5\) (Guimarães, 2008, 79). According to the author, the three trends differ in aspects of their theory and methodology and were developed in different places, as shown below.

“Subject cataloguing” refers to the development of catalogues according to subjects, first developed in the United States of America, at the Library of Congress, where in 1876 Charles Ammi Cutter codified the rules establishing the formation and control of alphabetical subject headings.

“Indexing” developed from the British interest in the elaboration of subject indexes, which became increasingly more structured, particularly due to specialized information services, especially the publication of periodicals, which culminated in the middle years of the 19\(^{th}\) century, at Yale University, with the publishing of William Frederick Poole’s index, named “An alphabetical index to subjects treated in the reviews and other periodicals”. This trend was later enhanced by the studies of the Classification Research Group (CRG), also in England, in the 1950s and 60s, which conducted research based on Ranganathan’s classification.

The third and youngest thematic trend in KO, “analyse documentaire”, was developed by the French, from the perspective of documentation, in the context of research that involved information recovery methodologies and tools, with the support of linguistics. Marcel Van Dijk, Maurice Coyaud, Jacques Chaumier, Jean-Claude Gardin, are important names in this area. As an example of this trend we highlight the “syntagmatic organization language” (SYNTOL), built in the 1960s, by Gardin and his team. Being the principal of the Centre de Recherches Archéologiques (CNRS), Gardin aimed to develop the method of systematization to analyze and classify scientific papers more strictly. The approach followed by this trend was well characterized, especially in Spain, by many researchers, such as: Félix Sagredo Fernández, José Maria Izquierdo Arroyo, Antonio Luis García Gutiérrez, María Pinto Molina, Mónica Izquierdo Alonso, among others.

Another important issue is that the trends outlined with respect to ordering KO each display characteristics that, according to Guimarães (2003, 2008), involve moments of, respectively, art, technique and methodology construction. The first moment was guided by “um talento especial, uma verdadeira habilidade artística em que o emprego do bom senso se aplica a um processo altamente intuitivo: a determinação do conteúdo do documento e sua consequente nomeação”\(^6\) (Guimarães 2003, 104).

The second trend, in the 19\(^{th}\) century, focused on techniques and tools developed

---

5 “[...] nowadays there is a significant interlocution space in the scope of the International Society for Knowledge Organization (ISKO)”.

6 “a special gift, a true artistic skill in which the use of common sense is applied to a highly intuitive process: the determination of a document’s content and its consequent designation”.

specifically to address the increasing capacity of bibliographic creation and mass publication, as the production of many identical papers, with no mistakes or alterations, became a reality following the creation and development of the press.

The third and last moment mentioned by the author occurred in the 1950s, with the creation of scientific bases, seeking, in particular, to develop strict methodologies.

We understand that the second and third trends highlighted the investigative practice of KO, in relation to TTI. Despite the theoretical and practical complexity entailed by these trends in ordering information, it is a matter of common sense to make representations that enable effective recovery of a document’s thematic content. It is not only necessary to analyze subject cataloguing, in terms of its principles and respective subject headings, but also to study indexes as devices built on theoretical bases, and, in addition, to verify the principles, methodologies and tools of analyse documentaire, which involves both formal and subject content analysis, or depending on the approach, only thematic content analysis.

Graduate teaching in Brazil

In Brazil, the first four master’s graduate courses in library science were created in the 1970s. At Universidade Federal de Minas Gerais (UFMG), the course was named Library Administration and started in 1976, having two main branches: library science and education; and library Science and specialized information (Vieira 1990).

The master’s course at Pontifícia Universidade Católica de Campinas (PUCCAMP) began in 1977 and had one main area: library science teaching methodology (Lima, 2008).

At Universidade de Brasília (UnB), the master’s course started in 1964, but was postponed at the end of 1965 (due to the political situation in the country), and only effectively restarted in 1978, with two main branches: planning, organization, information systems administration; and resources and documentation techniques and scientific information (Mueller 1985).

The master’s course at Universidade Federal da Paraíba (UFPB), which began in 1978, offered public library systems (Lima 2008).

In 1986, the Escola de Comunicação e Artes (ECA), of Universidade de São Paulo (USP), created the master’s in Information Science and Documentation as a main area together with the Communication Science Graduate Program (Mueller 1985).

However, as mentioned before, the first master’s course in information science in the country was created in 1972, at a research institute, the IBBD. It was originated as a specialization course in information and documentation, from 1955, mentioned above. It had two main branches: information systems administration; and information Transference (Lima 2008).

The indicated courses were renamed “Information Science”, except the course at
PUCCAMP, which remained unaccredited during the three-year evaluation period from 2004-2006. More courses were opened in universities all over the country and, at ECA/USP, the graduate course became independent.

Considering the aforementioned three trends in thematic treatment of information, we can perceive their presence (albeit not in quantitative terms) at the three universities mentioned above.

In the 1980s, some papers in the context of subject cataloguing at UFMG can be found in the production of professor Marysia Malheiros Fiuza (1980, 1985, 1987), as Guimarães mentions (2008), probably leading to the current denomination of the undergraduate course program (Universidade Federal de Minas Gerais, online).

The indexing trend grew stronger in Rio de Janeiro, in the 1970s. This had much to do with the fact that following the creation of the master’s course in information science at the IBBD in 1972, and up until 1981, Frederick Wilfrid Lancaster was the foreign professor who supervised the largest number of essays, 34 in total (Robredo, Vilan Filho 2010, 196).

In addition, the trend in analyse documentaire started in Brazil with the research of the Temma Group, founded in 1986, at ECA/USP, due to the initiative of professor Johanna Smit. The publishing of the book Análise documentária: a análise da síntese (Smit 1987) is considered the official beginning of the group (Lara 2011). Moreover, Johanna Smit gained her doctorate in France under the supervision of Gardin, and her partnership with other teachers and researchers from USP has resulted in the expansion and strengthening of this direction of research which has taken place at Universidade Estadual Paulista (Unesp), campus de Marília (Guimarães 2011).

The adoption of the three trends proposed by Guimarães (2008, 2009) enables us to understand the fundamental context of TTI, in the field of KO, because it indicates the construction of the notion of subject. The examples mentioned are part of the path that characterize KO in Brazil, although we recognize that there are other courses and researchers in the country who have similarly embraced KO to a greater or lesser extent.

The precise nature of each of the three trends in question differ in relation to theoretical and methodological aspects regarding the notion of subject. Undergraduate courses in library science have been steadily growing stronger, while contributing to the establishment of one trend or another, as shown below.

Library Science undergraduate courses: the teachers and heads meetings of Mercosul

In Brazil, the first minimum curriculum for a course in library science was established in 1962, and in 1982, the new minimum curriculum was approved to substitute the previous one. In 1971, there were 17 library science courses in operation, and by 1977, this number had increased to 28 courses in the country (Mueller 1985).
Nowadays, the number of courses operating in Brazil has reached 40, as outlined below according to the adopted denominations: 30 bachelor degrees in “Library Science”; 2 bachelor degrees in “Library and Science Information”, 2 bachelor degrees in “Library Science and Documentation”; 1 bachelor degree in “Library Science and Sciences Information”; 1 bachelor degree in “Library Science and Information Unit Management”; 1 bachelor degree in “Library Science, enabling in Information Management”; 1 teaching degree in “Library Science” and 2 bachelor degrees in “Information Management” (Associação Brasileira de Educação em Ciência da Informação, online).

The Brazilian territorial expansion explains the quantity of courses being provided and their different names makes us suppose they have different conceptions, which needs to be verified. Thus, the notion of subject suggests processes and representation tools to apply to the thematic content of academic papers. In the search for conceptions and guidelines for pedagogical practices, there have been discussions between teachers and heads of courses. This movement was instigated 14 years after the second minimum curriculum.

Guimarães, Danuello and Menezes (2004) have said that from 1996 on, stimulated by the creation of Mercosul, a joint curriculum integration initiative was implemented in 43 undergraduate courses in library science from Argentina, Brazil, Chile, Paraguay and Uruguay. The inception of this project recalls the III Encuentro de Docentes e Investigadores em Bibliotecología, Ciências de la Información y Archivología (EDIBCIC), which took place in San Juan de Puerto Rico. These courses have defined seven areas for the undergraduate teaching, namely: Theoretical Foundation of Library and Information Science; Information Processing; Resources and Information Services; Information Technology; Information Units Management; and Professional Research and Practice.

Additionally, in 1996, the I Encontro de Diretores de Escolas de Biblioteconomia do Mercosul was held in Porto Alegre, which was a remarkable event in the history of Latin-American library science teaching, as it tried, for the first time, to discuss teaching issues. The process towards an integrated curriculum began through analysis and synthesis of the minimum content registered in the subject syllabus of the thematic areas (Guimarães, Danuello, Menezes 2004).

At the II Encuentro de Directores y I de Docentes de Cursos Superiores de Bibliotecología del Mercosur, which was held in Buenos Aires in 1997, in-depth discussion of the syllabus took place between the six working groups related to the previously defined areas (with the exception of the Professional Practice group, since its internal characteristics required institutional discussion) (Guimarães, Danuello, Menezes 2004).

At the abovementioned II Meeting, information processing – the main curriculum
area of this paper – had its name changed to “Information Organization and Treatment”. The participants in the discussion reported that it was difficult to work with contents stemming from different theoretical currents. Thus, Guimarães, Danuello and Menezes (2004, 2) show the necessity for this area to be disconnected from a conception focused on products and transformed into one that focuses “[...] com mais clareza [na] sua própria dimensão teórica, com especial contribuição dos estudos de Organização do Conhecimento (inclusive com inclusão explícita desse termo na ementa) [...]”7. Therefore, 1997 saw the first use of the designation “knowledge organization” in a formal teaching proposal for library science course content in Brazil.

However, only in 2000, at the IV Encuentro de Directores y III de Docentes de Escuelas de Bibliotecología del Mercosur, in Montevideo, the area “Information Organization and Treatment” presented reports from different schools, country by country, in relation to the methodological-theoretical issues in the field. Guimarães, Danuello and Menezes (2004) systematized 13 topics about the teaching of KO in Brazil, which has informed discussion of the teaching in the countries involved. The theoretical trends presented above result from the bibliography used by Guimarães and coworkers for the production of 13 topics, as follows:

1 Theoretical-conceptual aspects of subject information organization
2 Historical evolution of subject information organization
3 Theory of Knowledge Organization
4 Theory of the Concept (Dahlberg)
5 General Theory of Terminology (Wüster)
6 Classification in Archives, Libraries and Museums
7 Documentary reading
8 Documentary analysis
8.1 Concept identification
8.2 Concept selection
9 Documentary condensation (abstracts)
10 Documentary representation
10.1 Alphabetical documentary languages
10.1.1 Subject headings (Cutter)
10.1.2 Thesaurus
10.2 Notational documentary languages (bibliographic classifications)
10.2.1 Dewey’s decimal classification
10.2.2 Universal decimal classification
10.2.3 Faceted classification (Ranganathan and C.R.G.)
10.2.4 Library of Congress classification
10.2.5 Specialized classifications
11 Indexing systems and methods
12 Indexing politics
13 Information retrieval

In Brazil, from 2001 on, national curriculum guidelines for undergraduate courses

7 “[...] more clearly on its own theoretical dimension, with the special contribution of knowledge organization studies (and, moreover, with the explicit inclusion of this designation in the syllabus) [...]”.

were approved, aimed at leading to the development and implementation of political-pedagogical projects in university education institutions. As a result of continuing discussions, the minimum curriculum of library science was thus substituted by a more flexible curriculum based on these national guidelines.

Regarding the 13 topics listed at the Mercosul Meeting in 2000, about teaching in information organization and treatment, the notion of subject is linked to each of the denominations, as was expected. However, for an accurate analysis, it would be necessary to have the syllabus of each topic, so we would not suppose content for an expression that can have many meanings.

We observe in topic 4, the Dahlberg’s concept theory, whereby concepts are seen as knowledge units that connect to make up subjects. A similar relation can be established in topic 5, Wüster’s general theory of terminology, with regard to the element term-concept as a unit of meaning.

In topic 6, “Classification in Archives, Libraries and Museums”, we can debate the importance of the notion of subject in archiving and museology, and consequently question the use of the expression in library science, since the concept-theme “aboutness” and indeed other expressions widely used in the area have no consensual Portuguese translation.

Topics 7, “Documentary reading”, and 8, “Documentary analysis” indicate stages towards reaching the materialization of the document subject. Topic 9, “Documentary condensation (abstracts)” refers, initially, to a specific process and, subsequently to the product of this process (abstracts themselves), not forgetting to mention the difference between them, as they are presented as synonyms. The last stage, that of translation, is incorporated in topic 10, “Documentary representation”, in which many document tools of classification and indexing are indicated. The division displayed is done using codifications or words, respectively, notational or alphabetic documentary languages.

The last three topics, which are 11, 12 and 13, show systems aspects, methods and indexing politics, leading, lastly, to information retrieval.

The theoretical trends relating to TTI and how they systematize the 13 topics of teaching contents with respect to the same area, raises the following issues.

The division of documentary language types into formal aspects, as codifications and words, which is usual to Anglo-American approaches, is not adopted by the analyse documentaire trend. However, the adjective documentaire (documentary) as adopted to languages stems from that direction of thematic analysis. The term “documentary condensation”, given to the process of elaborating abstracts, is also a construction of the analyse documentaire trend. The subject cataloguing and indexing trends may not be so easily discriminated in Brazil, yet they are widely used all over the country as a reflection of the Anglo-American literature adopted here. Besides, the scientific and
curricular production makes the difference between the currents disappear, generally creating other differences.

Moreover, the expression “subject information organization”, in topics 1 and 2, seems to be a Brazilian proposal, developed from Foskett’s 1973 book and from other influences, since the term “classification” stopped being used to refer to the thematic aspects of KO.

**Final considerations**

The presentation of teaching approaches followed in Brazil allows us to observe basic elements about the notion of subject in KO, but it is only a partial view. This paper has not looked at other authors’ understanding about the composition of theoretical trends, nor has it given much attention to teaching matters. Even so, by highlighting one author’s point of view, namely that of Guimarães, we can get a clear picture of the various aspects regarding the notion of subject, of the theoretical trends in treating thematic elements and of Brazilian teaching.

The notion of subject must be studied more deeply so that the three theoretical trends mentioned can be compared, thus enabling better comprehension of and contribution to curriculum proposals, based on theoretical and methodological keystones, in place of aspects of teaching focused on the functioning of widely known tools.

The abovementioned studies should be continued, combining the work of different research groups, and more literature will no doubt be developed in this area, with the aim of understanding the prospects for teaching knowledge organization and the notion of subject, and thus enabling the planning of concrete actions to address the issues that must be solved.

If, on the one hand, understanding the theoretical bases that support knowledge organization and, moreover, understanding the notion of subject are necessary to thematic representation, on the other hand, their consolidation would allow a greater connection between undergraduates, graduates and research activities in library and information science.

**References**


Automatic indexing and ontologies: the consistency of research chronology and authoring in the context of Information Science

Abstract
An alternative to minimizing the semantic gap in automatic indexing systems is to make use of knowledge organization systems, like ontologies. Given that these two operational concepts are subjects of study in information science (IS), we consider it relevant to identify and analyze the continuity and consistency, chronological and authorial, between studies of ontologies and automatic indexing. We use as methodology an exploratory / descriptive study based on a systematic review. We conclude that a direct relationship (citation and co-authorship) between the articles analyzed is practically non-existent. Regarding exhaustiveness, specificity, precision and recall rates, we concluded that there are similarities among three of the bases used: LISS, LISTA and ISTA. LSD, while the fourth base used was the one that presented the lowest performance in retrieval, leading to low rates of exhaustiveness and specificity of the indexing process.

Introduction
Current society is no longer imaginable without considering access to information and content available in digital formats (Foskett 1997; Shera and Cleveland 1977). Digital content has been growing exponentially and is estimated to reach the ratio of five terabytes for each human being by 2020 (EMC 2017). In this context, information retrieval becomes an increasingly complex task that justifies the importance of using specific terms to describe and identify the content to be represented (Lancaster 2003; Stevens and Urban 1965).

In view of the big data phenomenon, dealing with large digital collections as currently available on the web has now become unfeasible using manual indexing. This situation has leveraged the use of automatic indexing, which dates back to the 1950s when there was the first spike in the availability of electronic texts (Baxendale 1958; Luhn 1957; Maron 1961). Even though information retrieval systems have evolved over the past decades, they are still not efficient, as one would expect by searching based on themes or concepts. Issues like polysemy and synonymy still hinder automatic indexing, and consequently document retrieval.

One alternative to minimize the semantic gap in automatic indexing systems is to make use of knowledge organization systems, like ontologies. Ontologies are defined here as engineering artifacts consisting of an intentional vocabulary used to describe a certain reality, along with explicit assumptions organized in a logical theory that represents concepts and relations in both clear and unambiguous ways (Gruber 1992; Guarino 1998). Although there is no consensus regarding typification of ontologies, three types often emerge: top-level ontologies; domain ontologies; and application
ontologies. The use of ontologies in automatic indexing systems, in particular, the use of domain ontologies, has already been explored. One can find these systems in an experimental stage, but presenting very promising results.

Given that these two operational concepts (automatic indexing and ontologies) are subjects of study with a long history in the field of information science (IS), the study of the relationship between them is still, to a certain extent, less known. The authors already took a step in this direction in a previous study (Simões, Machado, Souza and Lopes 2017), in which we used a corpus extracted from two databases (Library & Information Science Source (LISS) and Library and Information Science & Technology Abstracts (LISTA)). In that study we found many potentialities relating to ontologies and automatic indexing, many of them reported in the IS field. However, taking into account the authorship analysis (25 articles accounting for 72 different authors where only one is present in two papers), we consider it relevant to identify and analyse the continuity and consistency, chronological and authorial, between studies of ontologies and automatic indexing. In this context, we extended the study, by including two more databases (Information Science & Technology Abstracts (ISTA), and Library Science Database (LSD)). We, therefore, aim:

(i) to identify and account for scientific papers dealing with these topics (automatic indexing and ontologies) in the databases LISS, LISTA, ISTA, and LSD, and to explore their temporal distribution and departmental affiliations of their authors;
(ii) to verify the overlap of the articles retrieved in the databases mentioned;
(iii) and to assess the relationship between the studies by mapping the direct citations between them, and the co-authorship in the articles cited in the corpora.

We use as methodology an exploratory / descriptive study based on a systematic review.

Methodology
To meet our goals, we designed an exploratory and qualitative study based on a systematic literature review, along with content analysis techniques (Bardin 2011; Bernard and Ryan 2010; Gil 2008). To assess the relevance of the topics of automatic indexing and ontology, we adopted the same procedures as in the previous study (Simões et al. 2017). To collect the research the sample, we queried the four databases using the expression "index* AND ontolog*" in the fields title and subjects. And, to assess the thematic proximity, we worked with four categorical variables that were expressed on a negative scale of intensity (Bardin 2011), (see Table 1).
We gathered the collection on October 10, 2017, from the four databases listed. In the comparison between the four, we applied the formulas presented by Ribeiro (1996) to the rates of precision\(^1\) and recall\(^2\). The judgment of the relevance for the corpus is given by the criteria adopted and previously presented.

Concerning the comparison of the indexing terms observed in each database, we assessed the exhaustiveness and specificity of each database according to the following criteria:

- for the first concept, we computed the mean value of assigned terms;
- and, for the second, we considered the percentage of records that contained the terms ontology (s) and/or automatic indexing in the subject field, (half value if only one of the terms is present).

We want to note that, for the latter two concepts, we used the records of the relevant documents recovered in all four databases.

For the coincidence rate we used the Jaccard index \((| A \cap B | / | A \cup B |)\), which is expressed by dividing the cardinality of the intersections of the two groups by the cardinality of their unions.

Regarding the direct relations of citation and co-authorship between the studies that constitute the corpus, we used techniques for analysing social networks and co-occurrence (Alvarenga 1998; González-Teruel, González-Alcaide, Barrios and Abad-García 2015; Scott 1991; Sugimoto and McCain 2010).

**Results and discussion**

As a first result, we can say that the extension of the study to two other databases (ISTA and LSD) did not add new works to the corpus collected in the previous study (see table 2), so that their temporal distribution remains equal between the years 2003 and 2016 (see Figure 1).

---

1 The fraction of relevant instances among the retrieved instances.

2 The fraction of relevant instances that have been retrieved over the total amount of relevant instances.
Table 2: Corpus of the study

<table>
<thead>
<tr>
<th>Ref</th>
<th>Article</th>
<th>Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>02</td>
<td>Kabel, S. et al. (2004) <em>The added value of task and ontology-based markup for information...</em></td>
<td>-0.5</td>
</tr>
<tr>
<td>04</td>
<td>Ciravegna, F. et al. (2006) <em>Annotating document content: a knowledge-management...</em></td>
<td>0</td>
</tr>
<tr>
<td>05</td>
<td>Köhler, J. et al. (2006) <em>Ontology based text indexing and querying for the semantic web</em></td>
<td>0</td>
</tr>
<tr>
<td>06</td>
<td>Tsinaraki, C. et al. (2007) <em>Interoperability Support between MPEG-7/21 and OWL in...</em></td>
<td>0</td>
</tr>
<tr>
<td>07</td>
<td>Hernandez, N. et al. (2007) <em>Modeling context through domain ontologies</em></td>
<td>0</td>
</tr>
<tr>
<td>08</td>
<td>Pirrò, G. et al. (2008) <em>LOM: a linguistic ontology matcher based on information retrieval</em></td>
<td>-1.5</td>
</tr>
<tr>
<td>09</td>
<td>Allampalli-Nagaraj, G. et al. (2009) <em>Automatic semantic indexing of medical images using a...</em></td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>Solskinnsbakk, G. et al. (2010) <em>Combining ontological profiles with context in information...</em></td>
<td>-1.5</td>
</tr>
<tr>
<td>14</td>
<td>Kara, S. et al. (2012) <em>An ontology-based retrieval system using semantic indexing</em></td>
<td>-0.5</td>
</tr>
<tr>
<td>15</td>
<td>De Maio, C. et al. (2012) <em>Hierarchical web resources retrieval by exploiting Fuzzy Formal...</em></td>
<td>-1.5</td>
</tr>
<tr>
<td>16</td>
<td>Chiaraavalloti, M.T. et al. (2012) <em>URT &quot;Indexing and Classification Systems &quot;Projects and...</em></td>
<td>-1.5</td>
</tr>
<tr>
<td>17</td>
<td>Willis, C. et al. (2013) <em>A Random Walk on an Ontology: Using Thesaurus Structure for...</em></td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>Qiu, J. et al. (2014) <em>Constructing an information science resource ontology based on the...</em></td>
<td>-1.5</td>
</tr>
<tr>
<td>19</td>
<td>Thenmalar, S. et al. (2014) <em>Enhanced ontology-based indexing and searching</em></td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>Bendib, I. et al. (2014) <em>Semantic ontologies for multimedia indexing (SOMI)...</em></td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>Vlachidis, A.H. et al. (2016) <em>A Knowledge-Based Approach to Information Extraction for...</em></td>
<td>0</td>
</tr>
</tbody>
</table>

* Degree of thematic approximation determined by the sum of the value attributed to the proximity of the approach to each of the two concepts (see table 1), which must be greater than -3.

Figure 1: Temporal distribution of the constituent articles of the analysis corpus

Regarding the distribution of the affiliation of authors by area, we should note that the stated total (78 = 100%) is higher than the number of actual authors (72). This deviation originated from the fact that we considered the Department of Computer and Information Science (two occurrences) and the Department of Social Science Informatics (four) as both Computer Science (CS) and Information Science (IS). This decision set the difference of 11 percentage points in favour of CS (49%) in comparison to IS (38%), while the remaining 13% affiliated to other areas (Astronomy and Health...
Science, with 1 and 9 authors, respectively) (see figure 2, chart a)).

Figure 2: Departmental affiliation of authors

The difference between Computer Science and Information Science coverage disappears, however, when accounting is carried out on the number of papers (see figure 2, chart b). There is an equal distribution of IS and CS papers (eight for each, representing 64% of the total of 25). There are also five articles (20%) that combine the two and, finally, each one of the other papers represents a single remaining area or combination (4%, each).

On the comparison of the four databases, there is a significant difference between LSD and the other three (see Figure 3). The perceptual amplitude of the difference between the LSD and the other three bases vary in the precision rate (between 9 and 21 points) and in recall (between 26 and 40). In the indexing terms aspects, the difference is, for exhaustiveness 16 to 27 percentage points and for specificity, from 4 to 14. In the same matter, LSD values were affected by the existence of two registers with no assigned term (Ref 01 and 02), a situation that, in LISS and in LIST, occurs in another register (Ref 21). In this comparison, the LISTA base is the one that presents the greatest similarity in results; in terms of completeness and precision with the ISTA base and in terms of recall with the LISS.

Figure 3: Results of measurements (recovery and indexing)

<table>
<thead>
<tr>
<th></th>
<th>LISS</th>
<th>LISTA</th>
<th>ISTA</th>
<th>LSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieved documents (Rd)</td>
<td>31</td>
<td>28</td>
<td>17</td>
<td>9</td>
</tr>
<tr>
<td>Relevant Retrieved documents (RR)</td>
<td>20</td>
<td>21</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>Relevant documents (Rt)</td>
<td>21</td>
<td>22</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>Precision rate (RR/Rd)</td>
<td>65%</td>
<td>75%</td>
<td>76%</td>
<td>56%</td>
</tr>
<tr>
<td>Recall rate (RR/Rt)</td>
<td>95%</td>
<td>95%</td>
<td>81%</td>
<td>56%</td>
</tr>
<tr>
<td>Indexing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exhaustivity</td>
<td>77%</td>
<td>66%</td>
<td>66%</td>
<td>50%</td>
</tr>
<tr>
<td>Specificity</td>
<td>48%</td>
<td>41%</td>
<td>38%</td>
<td>33%</td>
</tr>
</tbody>
</table>
The percentage amplitude of the difference between the LSD and the other three databases vary in precision from 9 to 20 points, and in recall from 25 to 39.

The percentage amplitude of the difference between the LSD and the other three databases vary in precision from 9 to 20 points, and in recall from 25 to 39. Concerning term indexing aspects, the difference is, in exhaustiveness from 16 to 27 percentage points, and in specificity from 45 to 145. The computation of indexing aspects of LSD values was affected by the existence of two records with no assigned term (ref 01 and 02), a situation that, for LISS and LISTA, occurred in another record (ref 21). In this comparison table, the LISTA database is the one that presents similarity to others in the results, although not simultaneously; in completeness and precision with ISTA, and in recall with LISS.

The same pattern, evidencing a difference between LSD and the other three databases, is verifiable in coincidence rates. The three lowest values in the two measurements (Relevant Recovered ($RR)$ and Relevant Existing Documents ($Rt)$) report the overlap with the other three (see Figure 4). These values, below 40%, contrast with values above 50% for the other databases. We observed that the LISS database displayed the highest overlap rates in both measurements.

![Figure 4: Database overlap rates](image)

As far as the relationship between the constituent studies of the *corpus* is concerned, there are only two direct citations between them (Ref. 19 cites Ref. 05 and Ref. 14 cites Ref. 06). Regarding co-authorship within the universe considered (*corpus* and its references, that is, 689 works), we verified that the link between the authors of the articles in the *corpus* is limited to the group participating in the studies themselves (see Figure 5), justifying the low network density (0.04).
In the sociogram of figure 5, which represents the co-authorship relations, 24 unconnected elements are visible, corresponding to the 25 works of the corpus under study, since there is one author who presents two papers (ref 21 and 25). Within some of the elements, we can find authors with other works in common (depicted with lines whose assigned weight is greater than one).

Conclusions
Given the non-increase of the corpus under analysis, with respect to the previous study, the conclusions regarding temporal distribution are maintained: the joint investigation of the two concepts (ontologies and automatic indexing) is recent in the information science field (little more than a decade), its studies having increased in the last five years.

The analysis of the authors’ departmental affiliations confirms the previously mentioned inference regarding the great proximity and interchange between the fields of IS and CS in the study of these subjects. It will be possible, as regards affiliation, to indicate a relationship between these operational concepts and the health sciences.

As far as the comparison between the four databases is concerned, we can conclude that there is a close proximity between the LISS, LISTA and ISTA databases, in terms of completeness, specificity, precision and recall rates. We can infer that in such a situation, it is closely related to the fact that these three bases are included in the EBSCO platform, while the fourth base, the LSD, is inserted in another platform, ProQuest. This last base was the one that presented the lowest performance in retrieval for which, according to the analysis performed, the low rates of completeness and specificity of indexation were decisive.
The methodology used in the collection of the corpus emphasized the importance of indexing, particularly its exhaustiveness for a higher recall rate. Although the LISTA database shows a recall equal to the LISS and exhaustiveness 10 percentage points lower than this, this does not contradict the previous statement since LISTA's revocation rate increased due to the retrieval of the title and not the subject of two relevant papers in other bases. We can conclude from the sample that all four databases show little specificity in indexing, with rates below 50% due to lack of use (LISS and LISTA), or even total absence (ISTA and LSD) of terms from automatic indexing.

In the domain of the coincidence rates between the bases, we can infer that the association between the two services (EBSCO and ProQuest) is the reason why we have close values for the LISS, LISTA and ISTA bases, as well as a great difference of these rates between the three bases listed and LSD. A comparison with previous studies (Vinson and Welsh 2014) is impractical given the different methodologies adopted and bases used.

Finally, regarding the relationship between the constituent studies of the corpus under analysis, we conclude that there exists a direct relationship between them in terms of the two indicators (direct and co-authoring). The tiny number of direct citations (two) combined with the lack of other co-authorship between the authors represented in the corpus, in addition to the articles analyzed in the study, points to a weak exchange between the various actors.

This last conclusion and its inference poses some uncertainty in terms of authorial consistency in the study of these subjects in the field of IC. This uncertainty is somewhat mitigated by the finding of previous work by authors from the corpus, both individually and in co-authorship, in the bibliographical references analyzed. In contrast, it raises the question as to why those papers were not retrieved at the time of the search in the bases. These concerns suggest future lines of inquiry, in particular with regard to the mapping of indirect relations, not only in terms of citations, but also the extension of co-authorship to others, in addition to the 72, which may act as bridges or peak (Scott, 1991), increasing the density of the network.

References
Amos David, Nadine Ndjock

Big data, Knowledge Organization and decision making: opportunities and limits

Abstract
A concept that is currently attracting much interest in the field of information science is the concept of big data. This paper tries to answer some of the questions raised by big data such as how it relates to knowledge organization and decision making, and what its opportunities and limits are. Big data is defined as a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage, and process low-latency data. It also has one or more of the following characteristics: high volume, high velocity, or high variety. Big data comes from sensors, devices, video/audio, networks, log files, transactional applications, the web, and social media - much of it generated in real time and on a very large scale

1. Introduction
A concept that is currently attracting much interest in the field of information science is the concept of big data. Our aim here is to try and answer some of the questions raised by big data such as how it relates to knowledge organization and decision making, and what its opportunities and limits are. This paper has been mainly inspired by Challenges and Opportunities with Big Data: A white paper prepared for the Computing Community Consortium committee of the Computing Research Association (Agrawal et al. 2012) and by the theme of the biennial conference “Transition from Observation to Knowledge to Intelligence”, which was launched in 2014 and is now to be organized by the West Africa chapter of ISKO (David, A. and Uwadia, C. 2016).

1 https://www.ibm.com/analytics/hadoop/big-data-analytics, visited 17/12/2017
The Big Data pipeline is presented as in *Erro! A origem da referência não foi encontrada.*

Summarizing the challenges with Big Data, (Agrawal et al. 2012) states that

“Heterogeneity, scale, timeliness, complexity, and privacy problems with Big Data impede progress at all phases of the pipeline that can create value from data. The problems start right away during data acquisition, when the data tsunami requires us to make decisions, currently in an ad hoc manner, about what data to keep and what to discard, and how to store what we keep reliably with the right metadata. Much data today is not natively in structured format; for example, tweets and blogs are weakly structured pieces of text, while images and video are structured for storage and display, but not for semantic content and search: transforming such content into a structured format for later analysis is a major challenge. Data analysis is a clear bottleneck in many applications, both due to lack of scalability of the underlying algorithms and due to the complexity of the data that needs to be analyzed. Finally, presentation of the results and its interpretation by non-technical domain experts is crucial to extracting actionable knowledge”.

2. From data to intelligence

Intelligence is required to enhance strategic decisions. Intelligence begins with the observation of the environment and identifying decisional problems. Information technologies have made it possible to dematerialize all forms of objects and human activities. They have made it possible to “observe” human activities and environmental changes. Dematerialization technologies, coupled with sensor technologies\(^2\) have made it possible to acquire limitless volumes of data. Hence, it is now possible to observe and

\(^2\) The basis of Internet of Things.
store data on the environment.

The challenge, then, is in how to derive knowledge from the big volume of data acquired following the process of observation. Data obtained in this way will necessarily concern past events, or ongoing events. Therefore, the basic principle of deriving knowledge from data collected is the retrospective analysis of what must have been collected. This principle will help to understand the past.

Another challenge is how to anticipate the future, based on past events. This can be considered as deriving knowledge from experience – predictive analysis.

Concerning retrospective analysis and predictive analysis, “It seems that the knowledge of history can therefore actually guide the understanding of the present and guide future action. Not as exact science, as shown by the shortcomings of forecasting and the historical analogy, but as empirical rich meaning and teachings” (Paulhac 2007). Predictive analysis is defined as “a set of business intelligence (BI) technologies that uncovers relationships and patterns within large volumes of data that can be used to predict behavior and events. Unlike other BI technologies, predictive analytics is forward-looking, using past events to anticipate the future” (Nyce 2007).

Strategic decisions concern the future and they are long term with respect to time frame (Management Study Guide, 2016).

Competitive intelligence (CI) can be defined as a process as presented in David (2013). It can also be considered as a rational decision-making process. The following stages represent the process:

1) Identification of needs in the form of decisional problems to solve or stakes (threat, risk, danger),
2) Identification of the types of result,
3) Identification of the types of necessary information to obtain the result,
4) Identification of the relevant information sources,
5) Validation of the information sources,
6) Collection of information,
7) Validation of the information collected,
8) Processing of the information collected for the calculation of indicators,
9) Interpretation of the indicators,
10) Decision making for the resolution of the identified problem.

Stages (a), (i) and (j) are of importance. Indeed, CI is a global process where the orientation chosen in every stage will determine the type of decision to make.

This decision-making process constitutes the bridge between observation and intelligence. The product of the process is the derived decision to make which will determine the strategy to adopt – the intelligence.
2.1. Current approach to using information for decision making

In the watch approach to problem solving, information is first collected, then verified for validity before being analyzed for its possible application to solving the problem in hand. We believe that this approach is generally not appropriate since it produces a lot of noise – gathering information that will be discarded because of its not being applicable to the problem in hand.

Figure 2: CVI model for information use in decision making process

Based on the observation above (generating a lot of noise), we propose that the steps be inverted, starting with the identification of the possible use of information for solving the problem at hand, verify the relevance and validity of the information sources and then collect the relevant information. However, identifying the possible use of the information for solving the problem at hand cannot be precise unless the decisional problem is well identified and specified. This emphasizes the need of the first step (a) in CI process.

Figure 3: IVC model for information use in decision making process

This approach (IVC model) helps to solve the problem of what data to keep and what to discard in the acquisition/recording stage of the big data pipeline.

Having demonstrated the central role of information in decision-making process, we propose an extended information system to enhance the decision process through the functionalities of such a system.
2.2. Architecture of an extended information system

Figure 4 shows the important components of the architecture of the extended information system (IS) that we propose.

- **A-B-C**: The information to be managed by the system is determined by the set of decisional problems as specified by the decision maker. As in rational decision making and in CI process, this is the most important part of the system since it will determine the relevance of the information to be collected.

- **C-D**: After collection, the information very often needs processing in order for it to be integrated in the database (data warehouse) and prepared so that the results of the system can be visualized. Big data techniques offer a means of responding to the potentially huge volume of data that may be generated.

- **A-E-D**: These components correspond to how the end-user will extract a subset of information from the system. This corresponds to the matching operation in information retrieval systems. We were inspired by the cognitive approach in human learning: observation, elementary abstraction, symbolization and annotation. We have implemented these approaches in some systems that we developed.

- **I**: In terms of knowledge discovery and creation of intelligence, this component is the most important. We proposed various forms of retrospective analyses as well as various forms of visualization of results (charts, charts, maps) with the possibility of dynamically varying the indicators and the mode of visual presentation. We were inspired by the concept of the observatory.

- **G**: This corresponds to the interpretation and decision phase in the decision process steps. We believe that this phase should be carried out by the decision maker just as he determines and specifies the decision problem.

Figure 4: Architecture of an extended information system
2.3. The observatory and dashboard concept

Essentially, an observatory system is composed of sensors for collecting data. The data is fed into the observatory system after necessary transformation or restructuring. An observatory system provides various modes of information visualization. These modes of visualization are grouped into dashboards. Generally, the dashboard is composed of one or more screens, each subdivided into two or more zones.

Figure 5: Illustration of a dashboard with variator

Figure 5 shows a dashboard presented on a screen. The screen is composed of three zones, each with a specific form of data rendering for information visualization; in this case pie chart, table and map. The dashboard also provides a means of varying a specific
parameter (or indicator) and the rendering of information is dynamically adjusted.

The dashboard corresponds to the information visualization layer of the information system, where information visualization is defined as the study of (interactive) visual representations of abstract data to reinforce human cognition. The abstract data include both numerical and non-numerical data, such as text and geographic information.

3. Conclusion

Whether in cloud computing or in the Internet of Things, the fundamental question in information systems is to determine the use of information. This will determine which elements of the real world to choose (or observe) and how to represent them (visualization).

We believe that it is erroneous to think that big data can produce patterns and trends from a huge collection of data without (1) reflecting on end-uses of the information, particularly for solving decisional problems, and (2) specifying what data to collect. For example, if we are interested in collecting all data from Facebook interaction, it is necessary to know what we want to do with that collection, so that attributes of the interaction to collect can be identified. The approach currently employed in processing big data is illustrated in Figure 2 Error! A origem da referência não foi encontrada. The approach we are advocating is the one in Figure 3, whereby the decisional problem is first identified and specified before information is then collected based on the problem identified, thus enhancing the processing and use of the information.

Hadoop is convenient for huge data storage and access. However, there is still a lot to be done to make the framework efficient for information access and analysis.

Our proposals in terms of models and tools have been implemented in some systems. Added-value information through visualization has been applied in two applications.

The first application was developed for the PhD thesis of Dr. Nadine NDJOCK (Ndjock, 2017), which concerns enhancing decision-making process through the visualization of information using the concept of the observatory, as applied to the educational system in Cameroon. The decision-maker obtains the evolution of indicators, which enables him to guide a strategic decision such as the efficient management of personnel or adjustment of a training program.

The second application concerns the membership management system applied to ISKO (International Society for Knowledge Organization). This system is used to manage the members of the association. Not only can the executive board and chapter administrators create, access and modify member profiles, but they can also obtain added-value information through visualization tools that allow identification of development strategies.

Neither application involves a huge volume of data but using the observatory approach means that intelligence can readily be derived from observation. These two systems thus represent key stages in the big data pipeline.
References


Classification and Knowledge Organization Systems: ontologies and archival classification

Abstract
This paper intends to expose the existing relationship between the knowledge organization systems called domain ontologies, in the form of terminological ontologies, and classification, more specifically, archival classification. Initially, it was based on the ideas proposed by Madsen and Thomsen in the article *Ontologies vs. classification systems* (2009), in which the authors expose, in a precise way, how the elaboration of a terminological ontology prior to a classification system can make the system better, more complete and more user-friendly. The present work opted for a similar approach, but applied to archival classification systems, considering that such systems have peculiarities, such as the fact that their purpose is to serve a specific organization with its own administrative body and functions. We conclude that the prior creation of terminological ontologies could lead to improvements in the development of archival classification systems.

1. Introduction
Archival classification is one of the most important activities of an archivist, given that it is the basis for performing all other activities. However, it is challenging to achieve precisely because each institution has its own characteristics and administrative body.

In view of this, we have considered the potential relationship between a classification system and an ontology. The former seeks to organize a particular phenomenon, an organization in the case dealt with here, while the latter seeks to represent a given domain of knowledge. Due to the fact that an organizational environment can be treated as a domain, it can thus be represented through constructing an ontology.

Therefore, this study was based on an idea put forward by Madsen and Thomsen (2009) which suggests that developing a classification system with a terminological ontology as its basis is potentially a much better, more complete and user-friendly system, considering that a terminological ontology allows the archivist to easily visualize the entire functional structure of the organization, as well as being more flexible, enabling changes to be made whenever necessary.

In light of this situation, starting from the premise that it is possible to understand an organizational environment through a terminological ontology, that is, that the elements of an institution can be treated as concepts of a domain (the domain being the institution itself), it is also possible, therefore, to take this ontology as the basis for the construction of an appropriate classification system to deal with the documents produced in that institution.

In order to apply this idea, we constructed an ontology based on the existing elements in the organizational structure of an institution, the same as that on which archival
classifications are based: function, archive group, type of document, date, confidentiality etc. As a result, a clear picture emerged of all the attributions given to the departments, each of their functions, tasks and both intermediate and final activities, as well as the records used to achieve the proposed objectives.

This research thus offers a new perspective, linking studies of knowledge organization in general, focusing on ontologies and classification, with the field of archival science, which has several particularities, as will be discussed below. Moreover, it is important to point out that this research is based on other papers developed within the research context of knowledge organization that link archival science with knowledge organization systems such as Barros and Moraes (2012), Barros and Moraes (2013) and Barros and Tognoli (2015).

2. Archival classification

Classification within the archival science framework is a key activity as it unites management contexts with activities concerning access and record preservation. The essential principles of provenance and original order established at the end of the nineteenth century today remain fundamental to archival organization and classification, forming the basis for archival classification schemes.

Classification advanced in the early twentieth century notably following an analysis by Hilary Jenkinson published in 1922 which reconsidered what it entailed. For Jenkinson, classification is divided into two parts: “the institution, its history and organization and the second divides the files into classes and their subdivisions” (Jenkinson 1922, p 81). We can apprehend that both phases relate to contexts of records management in a scheme readily understood as a form of knowledge organization. We can explain archival science as a field that has an interdisciplinary relationship with knowledge organization, as defined by Hjørland (2008, p.86) "activities such as document description, indexing and classification performed in libraries, bibliographical databases, and archives [...]".

Archival classification, as a theory, developed fully in the 1950s with publications by Theodore R. Schellenberg, Oliver W. Holmes, Ernest Posner, Petter.

From the end of that decade onwards, the discipline made profound advances, but a division occurred that would be perpetuated over the years until the present day. This rupture concerned the separation of certain activities and concepts determined by which phase the documents are in.

One of the institutions responsible was the American National Archives, centered on the figure of Theodore R. Schellenberg, who spread and popularized his methods, separating classification as a typical activity of active archiving from the arrangement of permanent archives, thus initiating a systematic division in the field.

It is noted that no distinction had previously been made between these functions and the most commonly used term, "arrangement", was used for assigning things which
post-1950s American archival science considered to be different. Eastwood has undertaken a fundamental review of this terminological and conceptual problem, claiming that the choice of the term "arrange" to name the process is unfortunate since it implies placing things in an acceptable way, in a convenient order, such as arranging books on a bookshelf. The term classification is no longer appropriate, as in this case it implies the arrangement or ordering of things in classes and it is better used in archival science for the process of organizing active documents (Eastwood 2000, p. 93).

Today, classification by function is, at the same time, both the means and the basis to understand knowledge organization in the archival field. According to this perspective, classification is defined, by Shepherd and Yeo (2003), for example, as a system based on analysis of functions, processes and activities.

Classification is a managerial and planning activity, not only for archives, but also for administrations that generate documents. Thus, the construction of a classification scheme should include not only the relationship between records, but also the relationship between records and people, such that the classification system would become a useful facilitating tool of document description and appraisal. Thus, classification serves as the basis for every archival organization process and especially for archival knowledge representation.

With that in mind, recognising the importance of classification to the organization of archives, we established a relationship between archival classification and ontologies.

3. Ontologies and terms

Besides classification, there are several other tools for organizing information, some of which are still based on the concepts that form a given area of knowledge, among them ontologies. According to Madsen and Thomsen (2009), a classification system and an ontology can be differentiated by their purpose: a classification aims to subdivide a given phenomenon into classes in order to order it, whereas the ontology, which is a model, seeks to represent the knowledge regarding this phenomenon.

Ontology is a term that has different meanings depending on the field of study. In philosophy, it is a field of metaphysics that studies being, that which exists. In computer science and information science, it is a knowledge organization system based on the existing semantic relationships established between the concepts of a given domain of knowledge.

One of the most cited concepts of ontology is that of Gruber (1995), who states that “an ontology is an explicit specification of a conceptualization”. In a similar way, Borst (1997) defined an ontology as a “formal specification of a shared conceptualization”. Studer, Benjamins, and Fensel (1998) united the two concepts: “An ontology is a formal, explicit specification of a shared conceptualisation”. However, in spite of being very succinct, this definition ends up using some terms in a rather informal way, since
expressions such as “conceptualization” and “formal, explicit specification” carry with them a lot of meaning, as Guarino, Oberle and Staab (2009) point out. Thus, the definition proposed by Guarino (1998) may be of use, which states that “an ontology refers to an engineering artifact, constituted by a specific vocabulary used to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of the vocabulary words”.

In view of these definitions, we can understand that an ontology is an organization system, formed by a taxonomic structure of concepts referring to a given domain of knowledge. Among the classes and subclasses that form this structure there are a series of relations so that the organized domain can be understood in its various aspects. For example, Figure 1 shows a rather simple ontology representing a family.

**Figure 1: Family ontology made in Protégé.**

![Family ontology made in Protégé.](image)

In this example, we have the concepts “person”, “gender”, “daughter”, “father”, “mother”, and so on, as classes that form the ontology, considering that “daughter”, “father”, “mother”, “offspring”, “parent”, “sibling” and “son” are subclasses of the “person” class, since all beings present in these classes can be categorized as “person”, as distinct from what would occur supposing there were a “pet” subclass. Besides the classes, it is also important to mention the instances, which are the elements that form part of the classes. An example of this is readily apparent with respect to a particular family where we can say Bob Jones father of Ana Jones. Thus, Bob belongs to the “parent” and “father” classes, just as Ana belongs to the classes “offspring” and “daughter”. In this way, it is clear that the network of relationships could expand much further with the addition of other members of the family (which would also imply the creation of other classes, such as “grandmother”).

From the relations established between the classes of an ontology, an informational system can understand commands in natural language, even in cases of polysemy or homonymy, considering that the purpose of this organization system is precisely to
make clear any semantic relations existing between the terms of a given domain.

However, the present study worked with a specific type of ontology: terminological ontologies. Van Heijist, Schreiber and Wielinga (1997) provide the following definition: “Terminological ontologies such as lexicons, specify the terms that are used to represent knowledge in the domain of discourse”. This classification is given because of the structure of this type of ontology, considering that it is still a domain ontology, as explained previously. Madsen and Thomsen (2015) explain that a terminological ontology “allows multiple inheritance (polyhierarchy). Therefore [...], our terminological ontologies may be described as specific, domain specific, feature-based ontologies developed for concept clarification purposes”. Thus, it is clear that a terminological ontology has the objective of representing a knowledge domain considering the terminology used in it.

In addition, it should be mentioned that the present study worked with a semi-informal ontology (Uschold and Gruninger, 1996), that is, a terminological ontology developed in natural language, but still trying to maintain a defined structure, avoiding ambiguities.

4. Terminological ontologies as a basis for classification systems

This research is based on the hypothesis that a terminological ontology can facilitate the development of an archival classification system. Therefore, a semi-informal terminology ontology formed by general elements present in organizations was developed to demonstrate the construction of the classification system for the documents of an organization.

The study first attempted to take into account all the elements pertinent to an institution, or at least the most common representations of what an institution is. With this information in hand, it was possible to assemble the terminological ontology. As regards the ontology, developing the classification proved to be a simple task, since all the relevant elements were already properly organized, allowing others to be added in the future. The classification was made based on the hierarchical classification model, which allowed some flexibility in the organization of the elements.

4.1. The ontology of an organizational structure and the classification scheme

According to Sousa (2008), every organization has a reason for existence, called its mission, purpose or main objective. In order to reach their goal, organizations have an organizational structure, which is the set of elements responsible for carrying out the activities that the organization needs to do.

It is from the organizational structure that it is possible to visualize the division into departments, called organizational units, each one responsible for performing a certain function. There are several forms of departmentalization, but the one that uses functions as a criterion of division is the most common. To perform its function, a department
performs certain activities and these activities are formed by tasks. The tasks put into practice generate records, classified according to their contexts and provenance.

Thus, based on archival classification knowledge it was possible to develop the following ontology:

Figure 2: Ontology of organizational structure with examples of types of documents made in Protégé

Figure 2 presents the created terminological ontology. The boxes with a yellow circle are the classes of the ontology, formed by the structural elements of the organization, and the boxes with a purple lozenge are the instances, formed by the types of documents produced.

Looking at the ontology, it is easy to see that in this example there are three departments, each responsible for a specific function. In order to perform its functions, a certain number of activities are necessary, which produce specific records. It should be noted that some activities may produce similar yet different records, which should be considered in the archival classification system.

Thus, in possession of the ontology, the following classification system could be developed:
Table 1: Classification plan based on the terminological ontology.

<table>
<thead>
<tr>
<th>Macrofunction 1:</th>
<th>01</th>
</tr>
</thead>
<tbody>
<tr>
<td>Function 1:</td>
<td>01.01</td>
</tr>
<tr>
<td>Activity 1:</td>
<td>01.01.01</td>
</tr>
<tr>
<td>Records:</td>
<td></td>
</tr>
<tr>
<td>Bill</td>
<td>01.01.01.01</td>
</tr>
<tr>
<td>Activity 1.1:</td>
<td>01.01.01.02</td>
</tr>
<tr>
<td>Records:</td>
<td></td>
</tr>
<tr>
<td>Financial Statement</td>
<td>01.01.02.01</td>
</tr>
<tr>
<td>Activity 1.2:</td>
<td>01.01.01.03</td>
</tr>
<tr>
<td>Records:</td>
<td></td>
</tr>
<tr>
<td>Report</td>
<td>01.01.03.01</td>
</tr>
<tr>
<td>Activity 1.3:</td>
<td>01.01.01.04</td>
</tr>
<tr>
<td>Records:</td>
<td></td>
</tr>
<tr>
<td>Contract</td>
<td>01.01.04.01</td>
</tr>
<tr>
<td>License</td>
<td>01.01.04.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Macrofunction 2:</th>
<th>02</th>
</tr>
</thead>
<tbody>
<tr>
<td>Function 2:</td>
<td>02.01</td>
</tr>
<tr>
<td>Activity 2:</td>
<td>02.01.01</td>
</tr>
<tr>
<td>Records:</td>
<td></td>
</tr>
<tr>
<td>License</td>
<td>02.01.01.01</td>
</tr>
<tr>
<td>Activity 2.1:</td>
<td>02.01.01.02</td>
</tr>
<tr>
<td>Records:</td>
<td></td>
</tr>
<tr>
<td>Certificate</td>
<td>02.01.02.01</td>
</tr>
<tr>
<td>Activity 2.2:</td>
<td>02.01.01.03</td>
</tr>
<tr>
<td>Records:</td>
<td></td>
</tr>
<tr>
<td>License</td>
<td>02.01.03.01</td>
</tr>
<tr>
<td>Proposal</td>
<td>02.01.03.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Macrofunction 3:</th>
<th>03</th>
</tr>
</thead>
<tbody>
<tr>
<td>Function 3:</td>
<td>03.01</td>
</tr>
<tr>
<td>Activity 3:</td>
<td>03.01.01</td>
</tr>
<tr>
<td>Records:</td>
<td></td>
</tr>
<tr>
<td>Certificate</td>
<td>03.01.01.01</td>
</tr>
<tr>
<td>Proposal</td>
<td>03.01.01.02</td>
</tr>
<tr>
<td>Activity 3.1:</td>
<td>03.01.01.02</td>
</tr>
<tr>
<td>Records:</td>
<td></td>
</tr>
<tr>
<td>Financial Statement</td>
<td>03.01.02.01</td>
</tr>
</tbody>
</table>

As can be seen, from the ontology the development of the classification plan was simplified, and, in view of any potential need to alter the organizational structure, the ontology can easily be modified, serving as a guide for the modification of the classification. For example, it may be necessary for the organization to have one of its departments perform a new function, or it may be necessary to terminate the activities of one department. If such situations occur, the addition of a new function or the exclusion of one of the existing departments can be represented in the ontology in a very simplified way, allowing the archivist to perform the same modification in the classification system.

The classification plan shown in Table 1 is small compared to situations in large organizations, composed of many different departments, each performing several functions. The construction of the ontology prior to the classification system is useful
especially for such cases, since the quantity of documents produced is massive and having a tool to more effectively visualize the organizational structure can greatly improve the archivist’s work.

5. Conclusion

We have undertaken a theoretical and methodological exploration of a possible approach between archival classification and ontologies as a baseline, and demonstrated that using an ontology, an information representation tool, the construction of an archival classification system can be simplified, also providing completeness and flexibility to this system in a challenging and shifting world, in which most organizations are rapidly changing their structures and functions.

In real practice, the development of the ontology will differ completely from one organization to another, since each organization has its own characteristics. Its construction must be based on all aspects of the organization and on how to order the archival documents produced, which makes the construction of classification plans very demanding and difficult.

However, with this work, we were able to construct an example and a basis for this type of scheme, helping managers and archivists to construct better classifications.

As a fundamental principle, it is possible to work with more types of ontologies to help with the construction of classification plans in the archival field.

References


Ana Cristina de Albuquerque, Luciane de Fátima Beckman Cavalcante

Classification of photographs: methodological concepts in Archival Science, Library Science and Museum Science

Abstract
This abstract presents the methodological concepts involved in the classification of photographs in archival science, library science and museum science. A research project was conducted on Brazilian information science journals dating from 2007 and 2017, classified as A1, A2, B1 and B2 according to Qualis. Based on content analysis, it was clear that general studies of classification in the field of information science, should be deepened in order to provide professionals with more theoretical support, since the methodological principles discussed above were addressed in very few articles.

1. Introduction
The various types of documentation produced not only serve to prove, recall and witness facts, but they also use complex structures to reflect specific universes, condensed into characteristics and contradictions that will present a distinctive identity described by their keepers. The keepers will present the practices of their time and their relationships with the world and the people of that age by organizing the documentation or collection in a specific arrangement.

Organizing the volume of information that is available is a challenge. The organization of information and knowledge cannot be by chance: it is the result of all the transformations and influences, whether social, economic or cultural, that take place in the environment in which it circulates. The way it is ordered by its producers or keepers, as object of use outside a file, library or museum reflects the meaning of an era and of a culture. In this context, this article proposes a reflection on the classification of the photographic document in archival science, library science and museum science.

In view of the above, the objective of this article is to analyze methodological concepts of the classification of photographs in each proposed field. In the context of the research, the photograph is approached as a document full of peculiarities that have accompanied social changes and consolidated its numerous uses. The photographs collected during the lifetime of a person or an institution later become documents accumulated in institutions that also undergo transformations and social influences.

The mapping of possible methodologies used in each field was carried out through research on Brazilian journals related to the field of information science. The period studied was between the years 2007 and 2017, and the journals were analyzed were classified A1, A2, B1 and B2 according to Qualis. This analysis was undertaken using the content analysis method, which allowed us to categorize and discuss each point raised and systematized in the results.
2. Regarding the classification of photographs in Archival Science, Library Science and Museum Science

The term classification is characterized by the process of grouping and dividing knowledge according to its similarities, arranging the information so that its analogical relations stand out, such that the science, knowledge or documents can be captured in a precise way. Classification is a social phenomenon and in it, every fact drives the creation of new forms of classification between beings and knowledge. In this way, the actions presented in the area of social relations are classification acts.

The concept of classification in the fields of archival science, library science and museum science, matches philosophical concepts when they aim to give a theoretical basis to an activity that could be considered solely practical.

The concept of classification in archival science has at its core the principles of provenience and respect for original order. Due to the relationships between the documents themselves, and between the documents and their creator, archives have established an organic character, which, in accordance with these principles, demonstrates the importance of administrative relationships that unfold in certain ways to turn the document into an archival item.

Thus, classification is understood as an activity that will preserve the link between the hierarchical chains of the institution for easier retrieval of information, and demonstrate how those documents were constituted within each entity.

The circumstances in which photographic documents are treated in archives are well discussed in the literature and the effort to elaborate methodologies according to archival theory is visible even if it is not always carried out. For Heredia Herrera (1993) the treatment of photographic documentation occurs through its classification as soon as it becomes part of a collection and an index or thesaurus is used for retrieval. (Heredia Herrera 1993). The author also recommends that the selection must be made based on the elements and circumstances of its production.

For Lacerda (2008), the fact that there is a problem with the current techniques used in library science, causes archivists to mirror these techniques, which, in his view, results in the development of misguided approaches to photographic documentation to the extent that there is no adequate theory of the subject, on account of it being such a disputed area. The aforementioned author touches on a controversial point about the treatment of photographic documentation from an archival perspective. As regards classification of these documents, almost nothing is established and often the basis for this ends up being the treatment given to the materials by librarians. It is important to refer to Lopez (2010), since he assumes that one of the main differences between the documents of a library and a file is production. In a file the documents are produced in series by the producer and are not characterized by individual treatment, in this sense the author questions Lacerda’s (2008) contention that individualized and "special"
treatment is given to photographic documents. This separates it totally from classification activity.

According to Lopez (2010), when the image appears next to other text documents they can collectively be considered as a documentary unit, the image being associated with a certain set and thus already classified. The treatment then takes place through physical separation, due to the different types of conservation treatment (Lopez 2010, page 221). Lopez (2010) warns, in the same way as Heredia Herrera (1993), that any treatment should, first of all, come to be "classification-driven" (Lopez, 2010, 224). It is this that will guarantee the total contextualization of the documents that cannot be treated differently from the other documents of the archive (Lopez 2010, p. 224).

Understood in this way classification activity is intrinsically related to the final retrieval of the archival document, and the photographic document, like any other document in a file, needs correct contextualization so that its information can be retrieved. Within a permanent archive, these activities can sometimes be carried out with some adjustments since in practice a number of issues arise and hamper the work of specialists.

Bibliographical classifications were elaborated with the purpose of establishing the proper relationship between the documents, maintaining the intellectual characteristics of the existing thought systems developed up until the present. Classification systems have become essential instruments for the functions within an informational unit.

The ordering of documents according to the subject they deal with and their separation or junction, taking into account similarities and differences, is the basis of bibliographical classifications. "In fact, when we refer to the bibliographic classification, we mean a classification that has a basis of the matters dealt with in the documents" (Piedade 1983, p. 65). With photographic documents, the ordering elements, in the sense reported in this study, namely to classify its information, occur in the process of indexation, where information is taken from analysis and synthesis of its subject. Knowledge organization tools are also used to give access to documents such as controlled vocabularies, subject heading lists, thesauri and ontologies.

In a museum, what feeds and drives their existence are collections. Collecting has a fundamental role in the formation and management of a museum. To keep collections preserved and organized with the aim of showing them to the public, a place is needed where, besides the physical space, treatment of these objects is possible.

Respecting the traditions and practices of each field, the concept of classification is sometimes different, when conceived only from the point of view of the diversity of the collection and the storage of documents in each institution. However, it also reveals itself to be more consistent when seen from the point of view of information organization in each of these fields.
In relation to the documents of a museum, it is important to emphasize the curator's role in the acquisition and subsequent treatment of the photographic collections acquired. Carvalho and Lima (2000), speak of the importance of the curator's role in a collection with photographic collections. The authors write that, in the first place, the curator should reflect on how this collection will contribute to society, what its potential for knowledge is and how it will be used to generate new knowledge (Carvalho and Lima 2000, p.19).

The curatorial activity in a museum demonstrates the results of the studies in relation to the collection, providing an understanding of efforts undertaken to ‘organize’ iconographic documents, in every sense of the word. Lima and Carvalho (2005) write that it is necessary to understand the "historical nature" of image production so that it results in a better refinement of documentation systems, in describing it in a way that leads to the understanding of a social process and its availability to account for research and scientific approaches to these practices. The activities of curating and research in a museum lead to the perception that care is always made to keep institutional objectives well delimited. This is reflected in the analysis of the fields in question and in the observation of the collections that constitute major collections. It can be observed that classification is implicit in the documentary activities carried out, noticeable from the moment the choice and the acquisition of the documents are made.

3. Methodological concepts in periodicals

Analysis of the Brazilian journals studied, within a ten-year period from 2007 to 2017, and classified by Qualis from A1, A2, B1 and B2, demonstrated that regarding classification of photographic documents, the subject is barely discussed and sometimes not even mentioned. In fact, although there are many articles related to photography, these articles come with an implicit classification as part of the treatment process that the documents have undergone.

<table>
<thead>
<tr>
<th>QUALIS CONCEPT</th>
<th>YEAR</th>
<th>TITLE</th>
<th>METHODOLOGICAL CONCEPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A1</td>
<td>2006 The archival arrangement as written: a reflection on the narrative in images from the Pedro Miranda Fund in the Public and Historical Archives of Ribeirão Preto</td>
<td>Deals with a specific background and its arrangement, considering photography as the main element of a narrative, that is, to the extent that the arrangement is an intellectual activity and it has a narrative of its own to organize and make these documents available</td>
</tr>
<tr>
<td>2</td>
<td>B1</td>
<td>2017 Public health photography in the archives: department of leprosy prophylaxis in the state of São Paulo</td>
<td>It proposes a study on the treatment of photographic collections from archival methodology, Diplomacy</td>
</tr>
<tr>
<td>3</td>
<td>B1</td>
<td>2017</td>
<td>Photography and its tentacles: possible interpretations in the universe of archives.</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>4</td>
<td>B1</td>
<td>2010</td>
<td>The informational representation of photographic collections: the reconstruction of the senses through the use of software.</td>
</tr>
<tr>
<td>5</td>
<td>B1</td>
<td>2016</td>
<td>The photographic collections of the psychiatric hospital of Miguel Bombarda,</td>
</tr>
<tr>
<td>6</td>
<td>B1</td>
<td>2016</td>
<td>#Impeachment or '#naovaitergolpe': an analysis of folksonomy in the indexing of photographic images in social networks of web 2.0</td>
</tr>
<tr>
<td>7</td>
<td>B1</td>
<td>2016</td>
<td>Vila Tiberius: a heritage of the city in transformation. Photographic Documentation and the organization of information from the collection of Jornal da Vila of Ribeirão Preto - the monthly memory of a city.</td>
</tr>
<tr>
<td>8</td>
<td>B1</td>
<td>2017</td>
<td>The decisive moment of Henri Cartier-Bresson and indexation: an exploratory study of methods of indexing photographs.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>9</td>
<td>B1</td>
<td>2017</td>
<td>The feasibility of the Sara Shatford methodology for the indexing of photographs: the photographic collection of the UFRN School of Music</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>It addresses the concepts of documentary analysis and indexation and expresses the differences between the indexation of textual and iconographic documents, highlighting the need to use methodologies that are more appropriate to the document's form and presenting methodologies adapted to the informational treatment of images.</td>
</tr>
</tbody>
</table>

Nine articles were found, one in Qualis A1 and the other seven in Qualis B1 magazines. Of these nine articles, six specifically speak of archival collections, two use methodologies better known in librarianship and one in digital archiving. None of the articles discuss the treatment or classification of photographs in museums.

Only the article "The archival arrangement as written: a reflection on the narrative in images from the Pedro Miranda Fund in the Public Archive and Historical of Ribeirão Preto" discusses forms of documentary organization, that is classification in permanent archives. It does this by using narrative and argumentative concepts from history of art authors, archives and history, demonstrating the interdisciplinary field into which the photographic document is placed.

An interesting fact is that the article "Public health photography in the archives: the department of prophylaxis of leprosy in the state of São Paulo", demonstrates the importance of the use of diplomacy in regards to photographic documentation. This article does not speak specifically of classification but demonstrates that Diplomacy is an indispensable method for documentary genesis to appear thus enabling classification to be performed.

The articles "The decisive instant of Henri Cartier-Bresson and indexing: an exploratory study of methods of indexing photographs" and "The viability of Sara Shatford's methodology for indexing photographs: the photographic collection of the School of Music of UFRN" use these methods of classification and indexing in the field of librarianship and show the tradition in the treatment of photographic documents from analysis of subject. Their contributions also promoted the use of classifications as a significant contribution to the definition of keywords and the coding of documents.

In the article "#Impeachment or #Naovaitergolpe", Folksonomy concepts are used to propose an organization through online photo tags. This study requires knowing and mastering the precepts of bibliographic classification, in terms of defining the area of knowledge and the definition of specific terms. It has been noticed that the articles from the periodicals chosen, most often, deal with classification implicitly, but their precepts provide the basis for the treatment and organization of documents that have a great many particularities.
4. Final considerations

It was considered that the organization and classification given to collections of photographs, whether these are collections of an institutional or personal nature, reflect the thought of a specific time. The ways in which documents are organized and classified must be maintained and respected by the institutions which collect them so as not to erase the traces of their organicity, reflected in the manner in which they were collected and in what was interchangeable at that meeting. This task is challenging and requires the research work of professionals involved in the management process of institutions that have photographic collections, because the more a classification connects to a given time, the less suitable it will be for another.

Classification in these fields is frequently debated in relation to theory, forms of execution, and systems. However, when we examine the study of specific documents such as non-textual ones, the literature does not provide sufficient basis for processes that are different from those that written documents have to go through.

Thus, classification is not limited to assigning numbers, codes and subdivisions to activities and documents, but rather is an intellectual organization process in which the characteristics, information and context of each document reflect, to a greater or lesser degree of complexity, the functions and activities developed by an institution, the life of a person or the facts pertaining to a city or country.

The main function of filing a document is the information about a given fact. It is clear that studies on archival classification, as well as general studies of classification in the field of information science, should be deepened in order to give more theoretical subsidies to professionals. This will be reflected when the user receives the document.

In librarianship the emphasis is on the processes that allow the retrieval of information. Instruments of knowledge organization enable research to focus on important concepts and contents that provide a basis for reflections on how to improve the treatment of these documents.

In museology, contextualization is crucial in following past "heritage" and trying to establish relationships with the functions and objectives of the institution, together with an understanding of how documentation can be used as a symbol of the development of an epoch and a given city. In addition, it can try to maintain unity and consistency within the current modes of organization and classification that permeate the entire process of documenting.

Photography is a technological means of creating and reproducing images in series. Because it is determined by these material conditions of production, it develops its own expressive language, that is, form and content are inseparably intertwined. For these reasons, photography presents unique conditions that determine its treatment in information units. So, whether in a file, in a library or in a museum, it will always present the same constituent traits and different functions will be given to information
in those collections. Depending on this function, its classification has an essential role for both the professional and the research user alike.

References


Lopes, Luis C. A Informação e os arquivos: teorias e práticas. Niterói: EDUFF; São Carlos: EDUFSCAR.


Edson Marchetti da Silva

Comparing the use of full text search between a conventional IR System and a DBMS

Abstract
Although database management systems (DBMS) and information retrieval systems (IRS) are typically used for problem solving in distinct domains, in the last decade they have been given internal functions allowing them to perform full text indexing and search by relevance from the best-match IR model. However, when searching for similar studies in the Capes Portal, with terms in both Portuguese and English, no related work was found in any of the scientific research databases. Within this context, this article aims to analyze the characteristics of implementation and use of full text indexing in a DBMS by comparing it with an IRS. The DBMS chosen to perform the experiments was Postgres 9.4 while the IRS selected was the Terrier IR Platform V. 3.6, developed and maintained by the School of Computing Science - University of Glasgow. The corpus used is composed of 1,260 scientific articles in PDF, which were first transformed into plain text using the Apache Foundation’s Tika application. This was done so that both tools worked with the same textual content. The objective was to evaluate comparatively the advantages and disadvantages of using these platforms which serve to help search for information contained in collections of textual documents. As a result, it was demonstrated that in the DBMS context, the use of keyword search in unstructured textual attributes can be used in an integrated way with other structured attributes, having the advantages of simplicity of use and efficiency in response time. This strategy can thus be better applied in information systems that work on such hybrid content.

1. Introduction
With the increasing proliferation of documents in digital format, it has become increasingly difficult to organize them by subject matter in order to facilitate a subsequent search. Borges (2009) outlines the evolution of indexing from when it was a manual process until the way it is performed automatically today. The author describes the main challenges in dealing with this issue. This difficulty arises from the intrinsic characteristic of documents, usually structured as cursive text and stored in binary files of the type doc, pdf, etc. In this context, textual contents are characterized by not having a predefined structure enabling them to be semantically represented so as to allow an exact search. The alternative is to process these documents by transforming them from binary to text format, and then to use word processing techniques to break them up into normalized terms so that they can be indexed in an inverted list. This structure enables a term-based search to be performed in order to find documents containing given terms among those in the collection. Such a search process is based on the application of several techniques, which enable a search to be performed that returns a list of documents ordered by relevance considering the best match of the search descriptors with those of the document collection. However, this does not mean that all returned documents are relevant to the user's search. In order to mitigate this
discrepancy, various techniques for dealing with the lexicon of terms to be indexed, and for ordering responses by relevance, have been proposed in research by several authors including: Lancaster and Warner (1994); Singhal et al. (2001); Manning, Raghavan and Shütze (2008) and Baeza-Yates and Ribeiro-Neto (2011).

In the last decade, leading database developers such as Oracle, MySQL, SQL Server, PostgreSQL, among others, have implemented data types that enable storage of documents in a customized way to optimize their retrieval by search terms; that is, these resources allow users to perform searches on data strings using features that go beyond the traditional search for regular expressions using the “like” and “rlike” commands inserted in the SQL query predicates. In this way, it is possible to keep both structured information and indexed textual information in the same DBMS. The demand for finding information from keywords has become prevalent in today's information systems where users want to retrieve information from textual contents, such as comments provided by other users, synopses of books, or even search for words in complete documents.

Although such DBMS-implemented resources extend the possibilities of retrieving information from document bases, when searching the Capes Journal Portal for terms in both Portuguese and English, there were no studies comparing the use of these platforms. Therefore, the idea is to carry out experiments that will identify some comparison indicators, such as: corpus load performance; search performance; ease of construction and customized solutions.

While this introduction has contextualized the theme, as well as describing the experiments, the rest of the paper is structured into six sections, the contents of which are as follows: Section 2 – objective; Section 3 – theoretical concepts; Section 4 – methodology; Section 5 – main findings; Section 6 – conclusions and recommendations for future works.

2. Objective

This study aims to compare two different means of retrieving information using full text search, comparing the advantages and disadvantages of the SGBD PostgreSQL and the Information Retrieval System (Terrier).

3. Theoretical concepts

Textual search operators have existed in DBMS for years. In the case of PostgreSQL, there are three separate approaches to pattern matching: the traditional SQL LIKE operator, the more recent SIMILAR TO operator and POSIX-style regular expressions. Besides other functionalities, the aim of all these operators is to identify whether determinate patterns occur in a string of text, that is, perform key word searches. However, they lack certain necessary requirements to be considered as modern information retrieval systems, such as: linguistic support, ranking of search results and
efficient response time. Therefore, a new functionality designated Full Text Indexing was implemented in version 8.3. According to Postgresql (2017), Full Text Searching provides the ability to deal with natural-language documents that satisfy a query, and optionally to sort them by relevance. This feature allows documents to be preprocessed, creating an index for subsequent rapid searching. Preprocessing includes parsing documents into tokens and converting tokens into normalized lexemes. Furthermore, the use of dictionaries allows fine-grained control over how tokens are normalized, so that it is possible to:

- define stop words that should not be indexed;
- map synonyms to a single word using Ispell;
- map phrases to a single word using a thesaurus;
- map different variations of a word to a canonical form using an Ispell dictionary;
- map different variations of a word to a canonical form using Snowball stemmer rules.

After preprocessing the tokens, they are stored in a tsvector datatype. The tsvector is a sorted list of distinct lexemes, which are words that have been normalized to merge different variants of the same word. Optionally, integer positions can be attached to lexemes. Optionally, integer positions can be attached to lexemes. The position indicates the word's location in the source document. Positional information can be used for proximity ranking. Position values can range from 1 to 16383. Figure 1 shows how Postgres transforms a cursive text to store it in a tsvector. The stop words are eliminated, the words are reduced by the stemmer algorithmics and put in lexicographic order followed by a colon and an integer numbers list that shows all the positions where the lexeme occurs in the source text.

Figure 1: Example of how the tsvector stores data

![Example of how the tsvector stores data](source.png)

According to Terrier (2017), the Terrier IR Platform is a highly flexible, efficient, and effective open source search engine, readily deployable on large-scale collections of documents. Terrier is an application written in Java that implements state-of-the-art indexing and retrieval functionalities, and provides an ideal platform for the rapid development and evaluation of large-scale retrieval applications.
As described by Apache (2017), Tika is a toolkit that detects and extracts metadata and text from over a thousand different file types (such as PPT, XLS, and PDF). All of these file types can be parsed through a single interface, making Tika useful for search engine indexing, content analysis, translation, and much more.

4 Methodology
This experiment was segmented in five steps described below.

4.1. Definition of corpus
In this step, a corpus was constituted with 1,260 scientific articles, in Portuguese and English, in PDF format, totalling 709 Mbytes.

4.2. Conversion of PDF files to text
To allow full indexing of the articles that make up the document collection, they must first be converted to text format. For this reason, the Apache Tika\textsuperscript{1} program was used, which receives a PDF file and converts it to text. Additionally, it was necessary to create and execute a parser in C++, to remove accented characters and text formatting characters by leaving the generated files in pure ASCII format. It should be noted that after the conversion of the data, the total file size was reduced to 47.7 Mbytes.

4.3. Loading data in PostgreSQL
In the third step, the textual contents were loaded in a table defined in the PostgreSQL V.9.4 DBMS. Since the text file loading process, executed through the PostgreSQL's Copy command, did not automatically convert the text content to the tsvector type, it was necessary to implement a "Before Insert" Stored Procedure and associate it with a "Trigger" before loading, as is shown in Figure 2. The total table load time was 8152 milliseconds. This load process indexes the documents into a word vector, generating a row in the table for each document.

Figure 2. Stored procedure/trigger implemented at DBMS

```
Create or replace function atuTsVector () returns trigger as $$
begin
    new.tsv_texto := to_tsvector(coalesce(new.txt_artigo,' '));
    return new;
end$$
LANGUAGE plpgsql;
Create trigger t_tsv before insert or update on artigo
for each row execute procedure atuTsVector();
```

Source: by the author

4.4. Indexing of text files by Terrier

\textsuperscript{1} Tika - https://tika.apache.org/.
In the fourth step, after completing installation of Terrier V.3.6, its desktop version was executed in order to index the same files in text format. The total indexing time was approximately 14.1 seconds.

4.5. Search for keywords

In the last step, some experiments were carried out in order to verify the behaviour of the responses in both cases.

5. Main findings

In this section we present the results of the comparison between the platforms studied. Table 1 shows information regarding the loading of the document collection.

Table 1: Information regarding the process of indexing documents

<table>
<thead>
<tr>
<th>Postgres</th>
<th>Number of files</th>
<th>Loading time (seg.)</th>
<th>Terms</th>
<th>Distinct terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1\textsuperscript{st} part</td>
<td>530</td>
<td>6.1</td>
<td>722,430</td>
<td>77,087</td>
</tr>
<tr>
<td>2\textsuperscript{nd} part</td>
<td>530</td>
<td>6</td>
<td>840,531</td>
<td>226,082</td>
</tr>
<tr>
<td>Postgres (1\textsuperscript{st} +2\textsuperscript{nd})</td>
<td>1,260</td>
<td>12.1</td>
<td>1,562,961</td>
<td>270,452</td>
</tr>
<tr>
<td>Terrier</td>
<td>1,260</td>
<td>14.1</td>
<td>1,502,815</td>
<td>177,325</td>
</tr>
</tbody>
</table>

Source: by the author

As can be seen in Table 1, Postgres indexes a much larger number of terms. This is because Postgres indexes each document individually; that is, it uses an attribute of tsvector type to store the distinct terms found in the document, each term being followed by a list with the positions where it was found. In this way, the same term can be indexed multiple times, one for each document in which it occurs. On the other hand, Terrier generates only one entry in the lexicon for each distinct term, considering all the documents in the collection. It is noted that after loading the first 530 files in Postgres, 722,430 terms were indexed, 77,087 of them being different. Similarly, after processing the second part with another 530 files 840,531 terms were indexed, 226,082 of which were distinct. It should be noted that these processes were performed in a separate database. Finally, a new database was created to perform the complete processing of the 1,260 files. As a result, 1,562,961 terms were indexed, of which 270,452 were distinct; that is, there was an intersection of 44,370 terms between the two systems. This is because each new set of indexed documents is very likely to have common terms. This huge difference between the number of indexed terms and distinct terms proves the different strategies adopted by both tools in the indexing process. In the case of Postgres, to optimize the search, it is optionally possible to create an index, Generalized Inverted Index (GIN) or a Generalized Search Tree (GiST) to a tsvector attribute existing in the table. In this way, it is easier to find answers during the search.
When comparing the total values produced by Postgres indexing with the Terrier-recorded values, it was verified that:

- Terrier time was 14.2% higher;
- the number of terms indexed in Postgres was 3.8% higher;
- The number of distinct terms indexed by Postgres was 34.4% higher.

As can be observed, Terrier reduced the number of distinct entries in the lexicon of indexed terms in relation to Postgres. This is due to a number of reasons: Terrier delimits the size in 20 bytes of the terms to be indexed; Terrier uses a more restrictive list of stop words with 733 items; Terrier restricts the size of the numerical terms to be indexed; that is, it looks for ways to reduce terms that tend to be irrelevant to the search.

Table 2 shows that both platforms returned the same set of responses, the difference being in ordering by relevance. The numbers in the table show the order that each technique put the documents into the response list. The Terrier column was used as reference for the ordering, with the aim comparing it with the answer from the six different Postgres ways. It is verified that there was no answer with the same ordering as the Terrier. The differences occur due to techniques used in each one of the tools.

<table>
<thead>
<tr>
<th>Document</th>
<th>Terrier</th>
<th>Rank1</th>
<th>Rank2</th>
<th>Rank4</th>
<th>Rank8</th>
<th>Rank16</th>
<th>Rank32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Doc2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Doc3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Doc4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Doc5</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Doc6</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Doc7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>2</td>
<td>7</td>
<td>7</td>
<td>6</td>
</tr>
</tbody>
</table>

Source: by the author

Table 3 shows a strong discrepancy between some of the results. This is due to the use of the stemmer algorithm, which considers 'ment' to be a suffix that can be removed from terms in the English language. For example, Terrier created entries in English such as: governor; governer and Portuguese terms without using the stemmer. On the other hand, Postgres created entries such as: governing; governac; governmental; government. Therefore, different treatments in data normalization and indexing processes can produce different sets of responses.
Table 3: Search for similar terms

<table>
<thead>
<tr>
<th>Terms</th>
<th>Number of documents found</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Terrier</td>
</tr>
<tr>
<td>Governamentais</td>
<td>145</td>
</tr>
<tr>
<td>Governmental</td>
<td>79</td>
</tr>
<tr>
<td>Government</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>172</td>
</tr>
</tbody>
</table>

Source: by the author

6. Conclusions and future works

From the empirical data it is verified that the indexation time in Terrier was 14.2% greater. However, considering that it corresponds to 2 seconds for indexing 1,260 documents, this result cannot be considered a significant issue between the two platforms. In addition, when comparing the search response times, both are instantaneous, that is, less than one second. Therefore, the search time cannot be used as a comparison term.

The other topic to be analysed are the quality of the answers. This is a complex analysis, because information retrieval is an approximation algorithm, and there are many factors that can interfere in the set of responses. This is due to the rules adopted in the parser of token separation, in the process of radicalization of terms (Stemmer), in the treatment of hyphenation, in the treatment of field types such as numbers and dates, etc. In addition, the relevance calculation algorithms used by the platforms are different. In this respect, we can consider the fact that Terrier implements the BM25 technique to be a point in its favour, this being a technique that, according to Silva (2013), presents better results than the other sorting techniques implemented by Postgres.

In conclusion, it is not possible to point out which platform is better than the other. However, this work demonstrates that the use of full text indexing using a DBMS mechanism as a place of information persistence has developed sufficiently to allow the treatment of large textual volumes, which can be treated within object-relational database schemas. This crucially reveals that integrating conventional data systems with the textual information processing system is something that can be handled in a very simplified way using a DBMS platform. Therefore, to develop an information system that needs to simultaneously deal as much with structured data as with textual data, the DBMS can be used advantageously as a persistence environment. This experiment thus pointed out differences and alternatives, but it cannot be conclusive in regard to stratifying the tools. It was considered that both serve their own purpose each with its specific characteristics and limitations.

This work can be extended using a larger corpus and testing different DBMS and IRS with the aim of analysing the scalability of both approaches.
References


Decio Wey Berti, Jr., Gercina Lima, Benildes Maculan and Dagobert Soergel

Computer-assisted checking of conceptual relationships in a large thesaurus

Abstract
We describe a method to support quality control of relationship instances in a large thesaurus or other KOS, using the example of AGROVOC (~33K concepts and ~97K conceptual relationship instances), where manually checking each relationship instance is not feasible. Our method identifies relationship instances that should be checked manually; it can also shed light on problems with the definition of relationship types. We apply a simplified version of the linguistic concept of verb valency to the analysis of conceptual relationships, treating relationship types as verbs. We map each of the two concepts in a relationship instance to an entity type; the resulting entity type pair is a valency pattern, as in the following example:

Flavivirus < causes > yellow fever ▬► Valency pattern [microorganism, diseaseOrDisorder]

A relationship instance that use a valency pattern that is rare for the relationship type might be erroneous and should be checked by an editor. We describe our method in detail, how we associated concepts with the appropriate entity type (this information is not available for AGROVOC) and how we organized the data for analysis. Then we present some illustrative results.

1. Introduction. The problem

Relationships between concepts form the skeleton of thesauri and other Knowledge Organization Systems (KOS). They are of enormous importance for the use of KOS to support retrieval and as knowledge bases for artificial intelligence applications. Many thesauri use only very broad conceptual relationships, hierarchical (BT/NT) and associative (RT). There has long been a call for refined relationship types (Schmitz-Esser, 1999; Soergel et al. 2004). Large thesauri have tens of thousands of relationship instances, thus introducing refined relationships requires huge effort. Relationships are useful only if they are of high quality. This paper addresses the issue of quality control in the establishment of relationships between concepts given that in most cases checking all relationship instances manually would be prohibitively expensive.

We use AGROVOC as our test environment because it is a large thesaurus and uses refined relationships, the focus of this paper. AGROVOC has ~33K concepts, and ~97K conceptual relationship instances; of these, ~35K use BT; ~35K use NT; ~4K RT, and ~23K use refined relationship types listed in the AGROVOC Ontology.

2. Literature review and conceptual background

2.1. Literature review

This paper is in the general area of finding errors in large KOS automatically. We saw three approaches.

Approach 1 uses purely formal checks (unprintable characters) to more content-
related checks (duplicate preferred labels, missing scope notes, or issues in the relational structure). The qSKOS program by Mader 2017 is a good example.

**Approach 2** uses use statistical analysis of text corpora to find, for example, instances of problematic equivalence between a term E in English and a term P in Portuguese. Using an English corpus and a Portuguese corpus on the same topic, one can check the occurrence patterns of E and P in their respective corpus; if E and P mean the same, then the occurrence patterns should be similar (Nohama et al. 2012).

**Approach 3** analyzes relationship instances using the entity types (semantic types) of the concepts that are connected. Mougin and Bodenreiter 2008 analyze the consistency of relationships in the NCI Thesaurus with relationships in the UMLS Semantic Network, but they apply this method only to derive a global measure of consistency, not to find individual errors in relationship instances. Jiang, Solbrig, and Chute 2012 also use UMLS semantic types to find errors in a specific type of KOS, a list of common data elements (CDE), such as Dosage Unit of Measure Code, in medical records with their associated permissible values. All permissible values must belong to the same semantic type, otherwise there is an error. In the example, the permissible value capsule is an error; capsule is not a dosage unit. The first study uses a top-down approach, starting with the UMLS Semantic Network. The second study's method is similar to ours but in a very specific and simple context. We use a bottom-up approach that starts with analyzing relationship instances in AGROVOC to find atypical relationship instances that should be checked by an editor; we could not find prior work that uses this method.

### 2.2. Conceptual background

We apply a simplified version of the linguistic concept of verb valency (Perini 2015) to the analysis of conceptual relationships. Summarizing from Perini: A verb may occur in one or more grammatical constructions or syntactic-schematic schemata. Such a schema, also called a *valency pattern*, specifies the syntactic and semantic (or thematic) roles of a verb's complements and possibly the types of concepts that can fill these semantic roles. For example, consider the construction or valency pattern

```
[6] VSubj > Agent  V  NP > Patient  with example
[7] The cook  melted  the cheese
```

```
[ ] VSubj > Agent  V  NP > Recipient  NP > Theme
[ ] Jim  gave  his girlfriend  a cake
```

Linguists study the valency patterns associated with a verb in a text corpus. The set of all valency patterns associated with a verb is the verb's valency.

In our approach to the analysis of relationships in a thesaurus or other KOS, we treat the relationship types as verbs. Most KOS are restricted to binary relationships (often represented as RDF triples); so the constructions are Concept1 V Concept2. Each concept has a semantic role and a concept type or entity type. We use a much simplified
form of valency patterns, a pair of entity types, \([\text{entityType1}, \text{entityType2}]\).

3. Methods

3.1. Methods, general principle
Our method is based on checking for each relationship instance the entity types of the concepts connected by the relationship. In the simplified world of binary relationships, such a pair of entity types constitutes a valency pattern, see Table 1.

<table>
<thead>
<tr>
<th>Valency pattern</th>
<th>Relationship instance example</th>
</tr>
</thead>
<tbody>
<tr>
<td>([\text{namedPlaceOrLocation}, \text{namedPlaceOrLocation}])</td>
<td>Canada (&lt;\text{spatiallyIncludes}&gt;) Fraser River</td>
</tr>
<tr>
<td>([\text{microorganism}, \text{diseaseOrDisorder}])</td>
<td>Flavivirus (&lt;\text{causes}&gt;) yellow fever</td>
</tr>
</tbody>
</table>

For each relationship instance we derive a valency pattern by mapping each of the two concepts to an entity type. This enables two types of analysis:

1. For a relationship type, list the associated valency patterns by frequency. The occurrence of several frequent valency patterns suggests that the meaning of the relationship should be examined. Infrequent valency patterns suggest that the corresponding relationship instances should be examined; the relationship may be used incorrectly. This is the aspect we are focusing on in this paper (Table 2).

2. For a valency pattern, list all associated relationships by frequency. This will shed light on the interpretation of the relationship types and may suggest some realignment of the definition and use of relationship types (Table 3).

<table>
<thead>
<tr>
<th>Relationship type: (&lt;\text{spatiallyIncludes}&gt;)</th>
<th>Valency pattern</th>
<th>Relationship instance example</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 good</td>
<td>([\text{namedPlaceOrLoc}, \text{namedPlaceOrLoc}])</td>
<td>Canada (&lt;\text{spatiallyIncl}&gt;) Fraser River</td>
<td>537</td>
</tr>
<tr>
<td>2 bad</td>
<td>([\text{physiographic Feature}, \text{physiogrFeature}])</td>
<td>lowland (&lt;\text{spatiallyIncl}&gt;) valleys</td>
<td>3</td>
</tr>
<tr>
<td>3 bad</td>
<td>([\text{physiogrFeature}, \text{namedPlaceOrLoc}])</td>
<td>Boreal forests (&lt;\text{spatiallyIncl}&gt;) Arctic tundra</td>
<td>2</td>
</tr>
<tr>
<td>4 bad</td>
<td>([\text{namedPlaceOrLoc}, \text{organization}])</td>
<td>Canada (&lt;\text{spatiallyIncl}&gt;) IDRC (International Development Research Centre)</td>
<td>1</td>
</tr>
</tbody>
</table>

The example in row 1 clearly makes sense; the relationship \(<\text{spatiallyIncludes}>\) can exist only between individual physical objects; AGROVOC further restricts the use of \(<\text{spatiallyIncludes}>\) to entities of type namedPlaceOrLocation. We can now turn our attention to the relationship instances that use low-frequency valency patterns. To make
sense of row 2, the relationship could be re-interpreted to apply to universals as well, in the sense that each namedPlaceOrLocation that is of type lowland <spatiallyIncludes> a namedPlaceOrLocation of type valley, which is clearly not the case. Row 2 would be a good relationship instance only under the interpretation that a namedPlaceOrLocation that is of type lowland often or sometimes <spatiallyIncludes> a namedPlaceOrLocation of type valley. Similar considerations show that row 3 is not a good relationship instance; what the editor wanted to express is that the Arctic tundra is a Boreal forest. The relationship in row 4 does not work at all since IDRC is not a namedPlaceOrLocation; the IDRC building is, so Canada <spatiallyIncludes> IDRC building would be ok.

Table 3: Analysis of relationship types associated with a given valency pattern

<table>
<thead>
<tr>
<th>Valency pattern [namedPlaceOrLocation, namedPlaceOrLocation]</th>
<th>Relationship type</th>
<th>Relationship instance example</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>&lt;spatiallyIncludes&gt;</td>
<td>Canada &lt;spatiallyIncludes&gt; Fraser River</td>
<td>537</td>
</tr>
<tr>
<td>1b</td>
<td></td>
<td>Argentina &lt;spatiallyIncludes&gt; Falkland Islands</td>
<td></td>
</tr>
<tr>
<td>1c</td>
<td></td>
<td>tropical America &lt;spatiallyIncludes&gt; Brazil</td>
<td></td>
</tr>
<tr>
<td>2a</td>
<td>&lt;includes&gt;</td>
<td>Madagascar &lt;includes&gt; Mangoky River</td>
<td>46</td>
</tr>
<tr>
<td>2b</td>
<td></td>
<td>United Kingdom &lt;includes&gt; Falkland Islands</td>
<td></td>
</tr>
<tr>
<td>2c</td>
<td></td>
<td>USA &lt;includes&gt; Guam</td>
<td></td>
</tr>
<tr>
<td>2d</td>
<td></td>
<td>Latin America &lt;includes&gt; Central America</td>
<td></td>
</tr>
<tr>
<td>3a*</td>
<td>&lt;hasMember&gt;</td>
<td>Francophone Africa &lt;hasMember&gt; Mauritania</td>
<td>212</td>
</tr>
<tr>
<td>3b</td>
<td></td>
<td>Latin America &lt;hasMember&gt; Brazil</td>
<td></td>
</tr>
<tr>
<td>3c</td>
<td></td>
<td>OECD countries &lt;hasMember&gt; United Kingdom</td>
<td></td>
</tr>
<tr>
<td>3d*</td>
<td></td>
<td>Small Island Developing States &lt;hasMember&gt; Belize</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>&lt;hasPart&gt;</td>
<td>USA &lt;hasPart&gt; Puerto Rico</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 3: Analysis of relationship types associated with a given valency pattern

From this table one can make many interesting observations and identify problems; we list here just a few. Rows 2a and 2d should be <spatiallyIncludes>. It appears that one meaning of <includes> in this context refers to political inclusion. 3b should be <spatiallyIncludes>. In 3a, c, and d we have on the left side groups (or sets) of named places; there should be a separate entity type for these (perhaps an issue with our approximate method of entity types, see Section 3.2). <hasMember> then makes sense if one uses a very loose definition. In AGROVOC, <hasMember> is also used for membership in organizations, such as ASEAN, which is a much more formal relationship. As seen from the prevailing examples, row 4 should be <includes>.

3.2. Methods, implementation

We needed to assign entity types to AGROVOC concepts so we could determine the valency pattern used in a relationship instance, accomplished through Steps 1 – 3.

1. A computer program constructed a tree-structure hierarchy of AGROVOC
concepts starting from _AGROVOC Top Concepts_ following NT relationships down (Fig. 1).

2. Starting from the _Basic Formal Ontology (BFO)_ class hierarchy, we developed a hierarchy of entity types specifically for AGROVOC by examining the AGROVOC hierarchy. (Fig. 2).

3. We then manually assigned entity types to concepts by taking advantage of the hierarchy: We identified segments of the hierarchy (some large, some small) so that all or most concepts in the segment belonged to the same entity types, as can be seen from Fig. 1. This assignment is approximate and contains some errors.

4. We also arranged the relationship types used by AGROVOC into our own hierarchy to support analysis (Figure 3).

5. We mapped concepts to their entity types and created a massive table of relationship instances to facilitate analysis (Fig. 4).

Figure 1: AGROVOC hierarchy pieces.

Figure 2: Entity type hierarchy pieces

**entities [AGROVOC sense]**
- world [for us: namedPlaceOrLocation]
  - continents
  - Americas
  - North America
  - Canada.

**organisms**
- microorganisms
  - viruses
  - Flaviviridae
  - Flavivirus

**features**
- physiographic features
  - land cover
  - vegetation
  - forests
  - forest types (by species)
  - coniferous forests
  - Boreal forests

**phenomena**
- biological phenomena
  - disorders
  - diseases
  - physiological functions
  - respiration

<table>
<thead>
<tr>
<th>E_1 BFO:continuant</th>
</tr>
</thead>
<tbody>
<tr>
<td>E_1.1 BFO:independentContinuant</td>
</tr>
<tr>
<td>E_1.1.1 BFO:materialEntity</td>
</tr>
<tr>
<td>E_1.1.1.1 BFO:object</td>
</tr>
<tr>
<td>E_1.1.1.1.1 inanimateObject</td>
</tr>
<tr>
<td>E_1.1.1.1.2 animateObjectOrganism</td>
</tr>
<tr>
<td>E_1.1.1.1.3 bodyPart</td>
</tr>
</tbody>
</table>

| E_1.3 BFO:specificallyDependentContinuant |
| E_1.3.1 BFO:quality == property |
| E_1.3.2 BFO:realizableEntity |
| E_1.3.2.0 stateCondition |
| E_1.3.2.3 BFO:disposition |
| E_1.3.2.3.1 diseaseOrDisorder |
| E_1.3.2.3.2 BFO:function |

<table>
<thead>
<tr>
<th>E_2 BFO:occurrent</th>
</tr>
</thead>
<tbody>
<tr>
<td>E_2.1 BFO:processBroad</td>
</tr>
<tr>
<td>E_2.1.1 process</td>
</tr>
<tr>
<td>E_2.1.1.1 processHappening</td>
</tr>
<tr>
<td>E_2.1.1.2 OBI:plannedProcessOrActivity</td>
</tr>
<tr>
<td>E_2.1.1.2.1 activity</td>
</tr>
<tr>
<td>E_2.1.1.2.2 methodTechnique</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E_3 nonBFOEntities</th>
</tr>
</thead>
<tbody>
<tr>
<td>E_3.1 namedPlaceOrLocation</td>
</tr>
<tr>
<td>E_3.3 scientificScholarlyArea</td>
</tr>
<tr>
<td>E_3.4 workersProfessions</td>
</tr>
</tbody>
</table>

OBI = _Ontology of Biological Investigations_
The actual table also includes the hierarchically structured notations for the entity types and relationship types, making it easy to "aggregate up" in the analysis.

4. Illustrative results

This section continues Section 3.2 and further illustrates our method at work examining the relationship types <includes> and <surrounds> with emphasis on shedding light on the definition and general usage of these relationship types.

Table 4 shows the top valency patterns for the relationship type <includes>.

<table>
<thead>
<tr>
<th>Valency pattern</th>
<th>Example</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[chemSubstance, chemlSubst.] heavy metals &lt;includes&gt; mercury</td>
<td>219</td>
</tr>
<tr>
<td>2</td>
<td>[macroorganism, macroorganism] Decapoda &lt;includes&gt; crabs</td>
<td>193</td>
</tr>
<tr>
<td>3a, 3b, 3c</td>
<td>[activity, activity] sectoral planning &lt;includes&gt; agricult. planning home economics &lt;includes&gt; cooking risk management &lt;includes&gt; risk assessment</td>
<td>80</td>
</tr>
<tr>
<td>4</td>
<td>[otherMaterial, otherMaterial.] soil parent materials &lt;includes&gt; rock</td>
<td>65</td>
</tr>
<tr>
<td>5</td>
<td>[taxonProperty, macroorganism] spring crops &lt;includes&gt; Triticum</td>
<td>63</td>
</tr>
<tr>
<td>6</td>
<td>[object, object] farm equipment &lt;includes&gt; harvesters</td>
<td>56</td>
</tr>
<tr>
<td>7</td>
<td>[namedPlaceOrLoc., namedPl.OrLoc.] USA &lt;includes&gt; Guam</td>
<td>46</td>
</tr>
<tr>
<td>8</td>
<td>[economicSector , economicSector] agroindustry &lt;includes&gt; fertilizer industry</td>
<td>30</td>
</tr>
<tr>
<td>9</td>
<td>[bodyPart , bodyPart] olfactory organs &lt;includes&gt;nose</td>
<td>12</td>
</tr>
</tbody>
</table>
The relationship type <includes> has 1,759 relationships instances; it uses 30 valency patterns that occur 10 or more or times, accounting for 1,192 relationship instances, 20 that occur 5-9 times, accounting for 134 relationship instances and 300 that occur 1-4 times, accounting for 433 relationship instances. Some of this variety may be due to errors in our approximate entity type assignments. Here we examine the top 7 valency patterns and two more selected for illustration. In most of the rows in Table 4 the valency patterns the two entity types in the pair are the same. Most of the relationship instances use <includes> in the meaning of set inclusion or hierarchy, one of the meanings of "includes" in natural language. This interpretation applies also to row 5; taxonProperty refers to a set of Taxa that have the property. However, natural language "includes" has several meanings (name a member of a group or class; cover; encompass, includes as constituent or part), and this ambiguity carries into the use of <includes> in AGROVOC. Row 7 expresses that USA as a political entity has a part Guam. In row 3a agricultural planning is a type of sectoral planning, but in row 3b cooking is a constituent activity of home economics, not a type of home economics, and the same analysis applies to rows 3c and row 8. So perhaps <includes> should be split into two relationships, or the hierarchical <includes> should be replaces with <hasMember> and the constitutive <includes> with <hasPart>.

Table 5 shows valency patterns for the relationship type <surrounds>. This relationship is related to <spatiallyIncludes> (mostly used for namedPlaceOrLocation) and <hasPart> (used for bodyPart, among others), with an important distinction: The heart is not a part of the pericardium, but it is surrounded by or enclosed in the pericardium. Similarly, the Gulf of Thailand is not part of Thailand but (partially) surround by Thailand. For bodies of water that are completely included in a country, the relationship <spatiallyIncludes> is used.

<table>
<thead>
<tr>
<th>Valency pattern</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 [bodyPart, bodyPart]</td>
<td>pericardium &lt;surrounds&gt; heart, 15</td>
</tr>
<tr>
<td>2 [namedPlaceOrLocation, namedPlaceOrLocation]</td>
<td>Thailand &lt;surrounds&gt; Gulf of Thailand 4</td>
</tr>
<tr>
<td>3 [otherMaterial, otherMaterial]</td>
<td>coffee pulp &lt;surrounds&gt; coffee beans, 1</td>
</tr>
<tr>
<td>4 [otherMaterial,developmentalStageAgeGroup]</td>
<td>cocoon &lt;surrounds&gt; pupae, 1</td>
</tr>
<tr>
<td>5 [physiographicFeature, namedPlaceOrLocation,]</td>
<td>Boreal forests &lt;surrounds&gt; Arctic region, 1</td>
</tr>
</tbody>
</table>

This table give rise to some general observations to clarify definitions of entity types and relationship types. As with <spatiallyIncludes>, the concepts connected with <surrounds> should be individuals, as in row 2. To make this work for row 1, we need
to interpret this relationship as: In any individual organism, the pericardium in this organism surrounds the heart in this organism.

Row 3 shows a problem in our entity assignments; coffee pulp and coffee bean should be bodyPart. Similar considerations apply to row 4, but the entity type of pupa and similar concepts that refer to the whole body or a body part at a given stage of development need further thought.

Row 5 is an error.

Looking at our large data table, we could make many other observations. To facilitate analysis, we prepared two major arrangements:

1 sorted by relationship type, then valency pattern and
2 sorted by valency pattern and then by relationship type

One observation relevant here is this: \(<\text{spatiallyIncludes}>\) is used for namedPlaceOrLocation, \(<\text{hasPart}>\) is used for BodyPart (among others), \(<\text{surrounds}>\) is used for both. This does not appear entirely consistent.

5. Conclusion

Large thesauri and other KOS that are well-structured using fine-grained relationship types are useful for information retrieval and reasoning in knowledge-based systems. But such KOS are hard to check and clean to assure that relationship instances are correct and then to maintain that high quality. We set out to develop a method that makes quality control feasible by identifying relationship instances that should be checked by an editor. We have demonstrated through numerous examples that our method of applying a simplified version of the linguistic concept of verb valency shows great promise for realistic quality control for large thesauri and other KOS.

Note

To obtain the full data illustrated in Figures 1-4, including the data table with the AGROVOC relationship instances, contact ds@dsoergel.com.

References


The Contribution of Semiotics to Knowledge Organization for music information

Abstract
The primary goal of knowledge organization (KO) is to determine which concepts should be considered as representing a field of knowledge. From a Semiotic point of view, different kinds of music meaning (being genuinely a concept or not) correspond to ways in which music is understood by listeners, how they communicate about music, and consequently, the conceptual universe of music domain. The present study discusses the particularities of interpretant levels, which for Peirce are: emotional (feelings and emotions), energetic (particular experiences) and logical (convention). Emphasizing the emotional level, the discussion highlights four theoretical implications for music information KO: 1) The levels of meaning and the use of music information are parameters that need to be observed when mapping the field of music; 2) Concepts in the field of music should be analysed from the perspective of the semiosis process, as it is not possible to operate with isolated elements; 3) Emotional concepts in the field of music do not follow linguistic conventions; 4) The function of objects in forming emotional concepts is not to adjust meaning to reality. We conclude that it is now more essential than ever to return to concepts constructed from an unconventional perspective in order to analyse music KO. As the semiotic approach holds the necessary flexibility to connect music-related concepts, the non-expert user's participation is fundamental in uncovering music domain.

1. Introduction
The primary goal of knowledge organization (KO) is to determine which concepts should be considered as representing a field of knowledge (Smiraglia 2014, 26). For that, the central point of knowledge analysis and the development of conceptual relationships is that of conventional verbal language. However, in the field of music information, words that refer to different levels of meaning are frequently used by those in search of information, as it can be seen from several online music websites. Allison-Cassin (2016) highlights the importance of considering elements that she calls "music scenes" for describing music, including rehearsal and performance spaces, production methods and music consumption. For her, the characterisation of music scenes introduces contextual metadata that should be taken into account in bibliographic descriptions of music, with "contrasting data structures focused on the idea of 'work' versus models that attempt to recognize the 'assemblage-like reality of music in an everyday context’" (Allison-Cassin 2016, 182). Fuller et al. (2016) used a large study group to test the efficiency of seven different types of persona (hypothetical archetypes of user behaviour) in classifying user profiles, with the goal of better directing music services. User behaviour on listening has also an impact on other meaning possibilities that include social perceptions of music (Selfridge-Field 2006;  

1 For example: Musicovery, Last.fm, Spotify, Superplayer, among others.
Laplante and Downie 2006; Laplante 2010, 2011), usage recommendations (Hu, Downie and Ehmann 2006), and emotional scopes (Hu and Downie 2007; Liebman, Stone and White 2015; Pesek et al. 2014). The structural characteristics of music may even hold a different relevance depending on whether users are interested in listening to music or in producing it (Andersen and Knees 2016).

From a Semiotic point of view, all these music-related elements correspond to ways in which music is understood by listeners, the meaning it evokes, and consequently, the conceptual universe of music domain. This is a pragmatic view, once the meaning of signs (being concepts or not) are known throughout its functions in the real world, i.e., practical phenomenological experiences, no matter if they are “true” or “false” in relation to reality.

This paper presents the Peircean Semiotics as a theoretical framework for music information domain representation. According to Peirce, the semiosis process is based on three correlates: sign, object and interpretant. The present study emphasizes the interpretant levels, which for Peirce (CP² 5.475, CP 5.476) are: emotional (feelings and emotions), energetic (particular experiences) and logical (convention). The objective of this study is to discuss the specificity of concepts resulting from each level of meaning, emphasizing the emotional level. After that, we present four theoretical contributions based on the Peircean Semiotics for knowledge organization for music information.

2. Levels of meaning

Within Peircean semiotics, the logical interpretant is based on social conventions, holding more stability than the other two levels. This kind of interpretation is a result of a deliberate analysis of the sign, similar to a logical structure of an argument. If a piece of music is in the "major A tune", it will always carry this characteristic, regardless of the way that users relate to this information, once this interpretation is a kind of a logical result of the analysis of music features and its relation according to previous rules. However, if users have an awareness of musical keys, they are likely to notice the "major A" key. The fact that music stimulates feelings of joy, sadness, celebration etc. does not alter the meaning, which is a general concept built on convention (Peirce 1995). The energetic interpretant in turn, is connected with the physical consequences that music causes as meaning effects, such as music to listen in the gym, or even the way emotion transcends itself and triggers physical responses such as crying, singing, dancing, etc.

On the emotional and energetic level, music has no rules as to how it should be felt, or should be interpreted (Cumming 2000). From this perspective, there is a theoretical gap in the field of KO that underlies the specificity of music related concepts. Concepts

---

2 The initial CP refers to the work "Collected Papers of Charles Sanders Peirce", and the following numbers refer to volume and paragraph, respectively.
from each level of sign interpretation have particular characteristics concerning the object-interpretant relation, and the extent of meaning sharing. On KO literature, it is possible to find some definitions of what is a concept, which vary depending on theoretical background. According to Dahlberg (1978) a concept is a set of true statements about an object. Therefore, it implies a dichotomy regarding the “true/false” relation, in which not all possible meanings could be considered. For Thellefsen (2004), a field must be analysed on its epistemological principles, functioning as a structure for developing a “knowledge profile”. That knowledge profile allows concept recognition, its analytical definition and relations. Thellefsen’s approach seems to be fruitful for scientific fields, in that its actors use theoretical bases, share objectives and research traditions. According to Hjørland (2007, 2) “concepts are the meaning behind words” and “one concept expresses one meaning”. In fact, all concepts are representations, but not all representations are concepts, as it could be a sensation, an image, a movement, etc. However, people need to communicate through concepts, and those meanings that are not genuinely concepts need to be “translated” into words. There is a distance between feeling an emotion or a physical sensation and naming it verbally. That is because some phenomenological experiences cannot be captured by linguistics terms. So, it is necessary a kind of adaptation of the words to express emotions, in this case “the meaning behind words” is a different kind of concept.

According to Savan (1981), emotion is composed of ill-defined feelings and sensations without clear identities, which makes them confused, yet similar. This is the location of the emotional interpretant, where the mind rehearses the understanding of the meaning and allows one to name emotions. Think, for example, of trying to analytically explain the elements that constitute the feeling of "joy": this feeling is not the result of a logical conclusion regarding the perception of a phenomenon (as it would occur in the analysis of a logical argument); rather, it is a hypothesis that this set of elements that come to mind should be called "joy". Savan (1981) calls this type of interpretant a "simplifying hypothesis", one which constitutes what the author calls “emotional concept”.

When music is used for recreational purposes, playing the role of sign, from the listeners’ points of view, emotional levels and personal experiences play a clearer role, revealing distinct concepts from those genuinely logic.

3. Methods
This paper is part of a wider study on knowledge organization for music information that has been developed along two main lines: one on theoretical semiotics, and the other one on applied semiotics. Barros, Café and Laplante (2016) presented some of the results of the applied study that involved interviews with young adults. The authors introduced music as a sign, and focussing on interpretants' emotional level, discussed the objects the music represents and its implications for KO.
We are now presenting results of the theoretical study based on Peirce’s works and his commentators. The focus relies on “concept construction”, as it is through this that knowledge and information organization processes are constituted.

4. Results

We propose four theoretical contributions to music information KO, distributed over two main lines, as follows:

1. The general constitution of music domain from a KO perspective
   1.1 The levels of meaning and the use of music information are parameters that need to be observed when mapping the field of music.
   1.2 Concepts in the field of music should be analysed from the perspective of the semiosis process, as it is not possible to operate with isolated elements.

   According to Peirce (CP 5.475), sign interpretation can occur only at the emotional level, or evolve to the energetic or logical level. The evolution of one level to another is related to, among other conditions, the intensity of the emotion felt, the recurrent validation of particular experiences at an energetic level and convention at a logical level. For example, the likelihood that a defined and more intense emotion (e.g. “excitement”) can evolve to an energetic interpretant (e.g. “feeling like dancing”) is greater than that of a less intense feeling (e.g. “a good song”). Thus, certain relations start to arise between meaning levels and, consequently, between concepts that constitute these levels, creating a framework for music domain representation.

   Emotions and particular experiences evoke non-referential concepts, i.e., it is not the features of “excitement” itself that are represented in a given song, it is the emotion felt by the user that is projected onto the song. This is to say that expressions such as "music to study" or "music to work out " go beyond mapping users' informational behaviour and become potential terms to represent the field with conceptual relations to emotions, musical structures, etc. In this sense, analysing signs, objects or interpretants in isolation does not provide sufficient references for concept analysis.

   Within interpretant levels, specific categories of terms can be inductively constructed based on users’ experiences and intensity indication (i.e., through tags, descriptions, “likes”) increasing the reliability of domain representation. For example, in constructing ontologies used by tag recommendation systems such as those in Font et al. (2014) and Font, Serrà and Serra (2015), meaning levels can operate as a principle for tag-based classification, in which the user could give more information to enrich semantically his or her classification choice. This method can be extended to different meanings, as for example, music genre definition: when a song is tagged as "samba" and "jazz" one could point in an intensity scale "how much samba or jazz is the song".

2. The nature of emotional concept construction

2.1 Emotional concepts in the field of music do not follow linguistic conventions.
2.2 The function of objects in forming emotional concepts is not to adjust meaning to reality.

Emotional concepts seem to be those that are least discussed in the field of knowledge organization, precisely because of their elusive character. Information Science refocuses on the universe of verbal language and more specifically, written language, which uses conventionalised symbolisation. This means that the bibliographic universe is influential in maintaining the current paradigm of KO (Smiraglia 2014). In the case of music, when viewed outside of its scientific field, it is more difficult to understand which elements are actually regularly shared between individuals, and we can see this in Abrahamsen's definition (2003) of the field of music as being everything and anything that may be connected with it.

Emotional meanings, communicated through emotional concepts (Savan, 1981), are central to the conceptual representation of the music field, more so for non-expert listeners. Moreover, due to the nature of the objects they represent and the iconic character of music signs per se, the logical and conventional conjecture that determines these concepts remains incomplete. Convention is the result of a previously established rule of significance. It involves a degree of generality, capable of maintaining the meaning provided, and associated with the concept in different situations (CP 2.292). Emotional concepts do not have previously established meanings, as what constitutes the interpretant in this case is not convention, but rather, the circumstances that suggest interpretations of the object's emanation in the sign. There is always a hint or possibility of assigning meaning. For example, while an Eric Clapton song can be always understood as such, thanks to the generalised conventional idea of “authorship”, we cannot expect that the same will occur with a “happy song”. On an emotional level, the same song could be understood as “happy” and “sad” at the same time, and this is not necessarily a contradiction. The object represented by a certain song can be different in each semiosis, and it is different from scientific concepts which need to be adjusted by the sign’s object to better represent it.

According to Cumming (2010) it is important to avoid expressions such as “X is a sign of Y”, in the sense that a musical sign interpretation must preserve the possibility for other suggestions. So, it is preferable to say that "certain characteristics of Y can be heard in X". In other words, we will not find an assertive answer about what are the characteristics of a certain emotion that also occurs in a given song. This dynamicity is intrinsic in music and, for KO purposes, it needs to be maintained in the domain representation throughout user’s constant contributions.
5. Conclusion

From a pragmatic point of view, we highlight the value of a semiotic theory that goes beyond the analysis of verbal language conventions, and considers the semiosis process to explain the music information domain. Anyway, theoretical deepening and analytical refinement is still needed for possibly developing methodological guides for music KO purposes. As the semiotic approach holds the necessary flexibility to connect music-related concepts, the non-expert user's participation is fundamental on uncovering music domain.

Further investigations will be developed from information retrieval perspective. On one hand, terms used in music description should not be the same as those used in music retrieval. On the other hand, terms that may not be used for music retrieval can be relevant for music description, giving clues to the user to make music choices and to improve automatic recommendation systems.

References


Critical questions for big data approach in knowledge representation and organization

Abstract
This study supports the importance of critically examining the advantages of big data approaches to knowledge representation and organization and carefully considering the cultural, technological, scholarly, societal and ethical implications of data-driven methodologies. Taking into consideration the complexity of the processes of representation and organization of knowledge, we argue that data-driven approaches have little to no effect on eliminating limitations and biases in existing knowledge organization and discovery systems.

Introduction
We live in the age of big data, wherein the production and analysis of large-scale data in relation to the various interactions of people, objects and technologies have recently become an everyday enterprise. However, as yet consensus has not been reached on the conceptualization of the big data. boyd and Crawford (2012) define big data as a cultural, technological and scholarly phenomenon that rests on the interplay of the technology that maximizes the computational power to collect, analyze, process and use data; methods of analysis that draw on large data sets to identify and interpret economic, social, technical and legal patterns; and the mythology associated with big data approaches that is based on the belief that they provide insights into higher levels of intelligence and knowledge than were previously available (p. 663). Ekbia and associates (2015) identify three main categories of definitions of big data in the current literature distinguishing among those that are (i) product-oriented with a quantitative focus on the size of data; (ii) process-oriented with a focus on the processes involved in the collection, curation, and use of data; and (iii) cognition-oriented with a focus on the way human beings… can relate to data” (p. 1525). They further develop the existing understanding of big data by adding the social movement perspective, which reflects the socioeconomic, cultural and political issues underlying the big data approach.

Most of the cultural heritage is inaccessible in digital format (Borowiecki and Navarrete 2017), with only estimated 17% of the cultural heritage collections digitized (Stroeker and Vogels 2014), however, the large-scale data approach to knowledge representation and organization has increased rapidly in recent years. For instance, cultural and research content included in European memory institutions has been valued for over €27 billion (Borowiecki and Navarrete 2017). The data-driven approach has also seemed to be inclined to the technological determinist view, the belief that technology as an underlying force in human society “has taken on a life of its own, … advancing according to its own inner dynamic, … unrestrained by social arrangement,
culture and thought. (Technological determinism is) the belief that technology acts as an independent force in our lives” (Volti 1992, p. 233). However, understanding of social change requires a sense for historical continuity, cultural and geographical contexts, and contingency of social processes that shape, inform, and limit the social appropriation of technology and cultural development. Along these ideas, the large-scale data approach to knowledge representation and organization has raised significant questions as to whether data-driven access to the knowledge would facilitate and/or transform the use and accessibility of the knowledge organization systems; whether it would help us to understand humans’ knowledge representation, organization and discovery behavior; or whether it would usher in new forms of biases, limitations, and privacy incursions. For instance, Ibekwe-SanJuan and Bowker (2017), discussing the relevance of a universal bibliographic classification and thesaurus, argue that big data will not remove the need for human constructed systems. They also suggest a shift from purely universalist and top-down approaches to more descriptive, bottom-up approaches that could potentially include diverse viewpoints in the representation and organization of the knowledge.

This paper aims to further develop understanding of the big data approach, its implications, and its peculiarities in the domain of knowledge representation and organization.

The Big Data approach in Knowledge Representation and Organization

It comes as no surprise that the domain of knowledge has embraced data-driven inquiry to advance representation, organization and discovery of knowledge. Application of semantic technologies have allowed linkage of information across various systems and platforms, and brought a new dimension to representation and organization of knowledge. The big data approach in knowledge representation and organization is mainly manifested in the following three main areas: i) linked data, that is, the application of semantic technologies to connect digital content across various systems; ii) visualization, the application of visualization techniques to present and discover knowledge, and iii) user-generated content, the utilization of user created content, such as tags and reviews, to complement traditional representation and organization of resources.

Linked data studies (e.g., Mayr and Zeng 2017) are grounded on Berners-Lee and colleagues’ (2001) seminal work describing the prerequisites of the Semantic Web, that is, the use of the Resource Description Framework (RDF) to uniquely define digital resources to ensure their discovery and meaningful relationships in digital environments. For instance, Baca and Gill’s (2015) paper explains how semantic technologies have enabled the multilingual and cross-cultural representation of the Getty Art & Architecture Thesaurus (AAT), the Getty Thesaurus of Geographic Names (TGN) and the Union List of Artist Names (ULAN). But even though the linked data
approach allows us to discover and understand the significance of cultural heritage objects across diverse contexts and cultures, issues related to the methodology behind the selected data, its provenance, credibility, and accuracy as well as the overall maintenance of linked data require careful consideration. For instance, Pattuelli (2012) points out limitations of using DBpedia as source of data, including human errors, incorrect classifications and missing content.

Börner and colleagues (2007) contend that visualization of the information reveals the structure, dynamics, properties and peculiarities of knowledge organization systems. Along with application of visualization techniques to existing knowledge organization systems (e.g., Hook and Gantchev 2017), there have also been experimental studies examining how visualization could potentially enhance resource presentation and discovery. For instance, Lin and associates (2017) developed a metadata visual interface, based on metadata aggregation from various digital repositories that have been enhanced by automatically assigned Dewey Decimal Classification (DDC) system class numbers. In addition to the linked data problems that are applicable to data visualization, usability and users’ overall understanding of the visualized data are main concerns (Chen 2005). In this regard, Ekbia and colleagues (2015) argue that visualization of the large data sets involves a “great deal of translational work…, which renders the accuracy of the claims problematic” (p. 1532). The opaqueness of data behind the visualization is further complicated by the tendency to provide more visually pleasing or aesthetic results rather than accurate mapping of the data. In this regard, Ekbia and colleagues argue that even though the “tradeoff between a visualization’s informational accuracy and its aesthetic appeal is readily acknowledged by scholars, these tradeoffs are not necessarily self-evident and are not always disclosed to users of data visualizations” (p. 1.533).

Among studies investigating how user-generated content can complement and enrich traditional representation of resources, many examine whether user-generated vocabularies (or tags) constitute indexing languages that can be utilized to represent the intellectual content of resources and/or how user-generated vocabularies are comparable to traditional systems of representation and organization, such as Library of Congress Subject Headings (LCSH) (e.g., Yi and Chan 2009). For instance, Ding and colleagues’ (2010) investigation of approximately 12 million tags collected from Delicious, Flickr and YouTube found consensus on 85% of the top 20 Delicious tags assigned in 2006 and 2007, suggesting the emergence of a shared social vocabulary. In contrast, George and colleagues’ (2008) longitudinal study of the stability of 1,804,379 Delicious tags assigned to 7,845 unique resources suggest that user-generated tagging vocabularies range widely from transitory and disparate to convergence on a shared vocabulary. These contradictory findings are not surprising as user-generated vocabularies are dynamic and reflect the diversity and richness of various contexts and
cultures. Also, besides issues related to data quality, utilization and interpretation of user-generated content raise ethical concerns, in particular, issues related to the users’ privacy as generally users are usually not asked to give consent or even informed that their online behavior or their digital data are being used in research. In this regard, boyd and Crawford (2012) argues that:

[user-generated] data were created in highly context-sensitive spaces, and it is entirely possible that some users would not give permission for their data to be used elsewhere. Many are not aware of the multiplicity of agents and algorithms currently gathering and storing their data for future use. Researchers are rarely in a user’s imagined audience. Users are not necessarily aware of all the multiple uses, profits, and other gains that come from information they have posted (p. 673).

Implications of Big Data approach in Knowledge Representation and Organization

Even though the big data approach has potential to augment the knowledge representation and organization, the big data “dual character,” that is, empowerment yet intrusion and constraints that put the “tensions and frictions of modernity in high relief” (Ekbia et al. 2015, p. 1540), needs careful consideration. In particular, it is important to carefully consider issues associated with the application of big data, such as fundamental questions about what we know about the world and about ourselves (epistemology), how we gain (methodology) and present (aesthetics) the knowledge, what techniques and technologies are advanced (technology) for these purposes, how these developments impact the nature of privacy (ethics) and intellectual property (law), and related social and political implications (political economy) (Ekbia et al. 2015, p. 1528). Along these lines, the data-driven approach to knowledge representation and organization raises the critical questions in relation to the quality of the utilized data, including its provenance, credibility and accuracy, that are warrant to the meaningful connection of vast amounts of digital data. As Boyd and Crawford (2012) have argued, “seeing patterns where none actually exists, simply because enormous quantities of data can offer connections that radiate in all directions” (p. 668) is potentially misleading and meaningless.

Despite the abundance of opportunities associated with the evolution of technologies, cultural heritage institutions also face difficulties to keep up with the technological developments to ensure accessibility of digitized collections. Moreover, digital cultural heritage collections based on a particular technology or platform might not be so easily converted to the newer version(s) and/or different platform(s). Even though the cloud technologies may serve as a time capsule to overcome these difficulties, availability of human and financial resources to maintain and keep up with the evolution of technology could be a challenge.

Another major concern is opacity of the algorithms behind the platforms and systems, such as Google search engine. Algorithms are defined as procedures for solving problems in a finite number of steps that frequently involve repetition of an
operation (Merriam Webster n.d.). Dourish (2017), discussing conceptual and technical differences between algorithms and programs, argues that while programs may implement algorithms, programs can be considered both more than algorithms because they include non-algorithmic elements and less than algorithms because of the constraints that are imposed by particular implementation requirements (p. 2). The prevalence of algorithms has been reportedly manifested in all forms of everyday interaction with information, from Google search engine’s suggested resources to Netflix’s recommended entertainment programs (e.g., Ananny and Crawford 2016; Gillespie, 2010). For instance, the library catalogue suggesting relevant and/or similar titles is perhaps one of the most common manifestations of algorithms in knowledge organization systems. Dourish (2017) asserts that understanding how algorithms work should be grounded on their relation(s) to other “computation forms, such as data structures” (p. 2), because they arise in practice and in relation to these forms. However, understanding how the underlying algorithms in knowledge organization systems work is hindered by their opaqueness. Application of semantic technologies that link user-generated content, such as book reviews, is another concern as it is unclear what methodologies and relevance criteria were utilized in selecting the links to the user-generated content (Hajibayova 2017).

Bias in knowledge organization systems is another major concern, as a number of studies have demonstrated the presence of biases in knowledge organization systems such as classification schemas (e.g., Dewey Decimal Classification, DDC) and controlled vocabularies (e.g., Library of Congress Subject Headings, LCSH) as well as biases of the knowledge organization specialists engaged in the process of representing and organizing knowledge. In series of studies Olson argued that controlled vocabularies tend to favor a mainstream view (e.g., Olson 2007), introducing “blatant biases or, more commonly, subtle, insidious marginalization” (Olson 2002, p. 6). Representation and organization of indigenous materials have also demonstrated a lack of understanding of indigenous epistemologies in the representation of indigenous cultures, resulting in limited and partial representation of indigenous knowledge (Doyle 2006; Metoyer and Doyle 2015). Algorithmic biases that are built into platforms and systems, such as the Google search engine and Wikipedia, are another major concern with regard to utilization of user-generated content. For example, Callahan and Herring (2011) found that quantitative and qualitative content analysis of Polish and English editions of Wikipedia revealed the systematic differences related to different cultures, histories and values with English language coverage having a significant advantage. Hinnosaar (2015) observed that women are less likely to contribute to Wikipedia, which leads to unequal coverage of topics and an overall large effect on contributions providing information about gender inequality.

While none of these concerns with the application of big data approaches to
knowledge representation and organization delineate the full scope of related issues, neither are they mutually exclusive, discussion of these issues in light of research on knowledge representation and organization can provide useful insights and serve as a solid starting point for critical analysis.

Conclusions
White, in his seminal work, wrote that technology “merely opens a door; it does not compel one to enter” (1962, p. 28). A central issue in the study of data-driven approach, should be the question of agency. In particular, investigation of extent to which data-driven approach facilitates the humans’ knowledge representation and discovery. Along these lines, this study suggests that it is necessary to critically interrogate the advantages of big data approaches to spark conversations about the cultural, technological, scholarly, societal and ethical implications of data-driven approaches to knowledge representation, organization and discovery. Taking into consideration the complexity of the processes of representation and organization of knowledge, we argue that data-driven approaches have little to no effect on eliminating limitations and biases in existing knowledge organization and discovery systems. Therefore, it is argued that, while data-driven approaches would be certainly valuable for the provision of a large-scale representation of knowledge, to ensure its meaningful and useful discovery a careful consideration of implications of these approaches is needed.

References


Giulia Crippa

Developing a field of knowledge through bibliography: art history in the 16th century

Abstract
We propose a discussion on two 16th century writings, specifically the fourth chapter of Giovan Paolo Lomazzo’s *Idea del tempio della pittura* (first published in 1591), dedicated to “the ancient and modern writers of art” (Lomazzo, 1971, p. 34) and the chapter “Painting precepts transmitted by the ancients and the moderns”, from Antonio Possevino’s *Tractatio de Poesi et pintura ethnica, humana et fabulosa collata cum vera, honesta et sacra* (1595). Alongside the much better-known Vasari, consideration should be given to Lomazzo and Possevino, since both authors began to elaborate systematic approaches to art books, specifically by compiling bibliographies, which was somewhat of a fresh means of expanding what was then a new field of knowledge, namely art. The present study uses a historical approach, focusing on original sources, so as to understand the choices in art literature that formed the basis and marked the frontiers of knowledge organization (KO) in the field of art as developed within art history. Source studies and literary reflection combine to inquire what kind of knowledge field was designed by these authors and how it determined developments in art studies.

1. Introduction: about the history of art history and the purpose of this paper
The purpose of this paper is to analyze two writings from the second half of the 16th century, written respectively by Giovan Paolo Lomazzo – the fourth chapter of his *Idea del tempio della pittura*, published in 1591 – and Antonio Possevino – “Painting precepts transmitted by the ancients and the moderns”, from his *Tractatio de Poesi et pintura ethnica, humana et fabulosa collata cum vera, honesta et sacra*, published in 1595. We wish to view these writings through the perspective of knowledge organization (KO). We recognize the need to increase critical study of KO in the field of art, considering that historical sources have traditionally tended to be read with a focus on how esthetical conception has evolved over time. Not much attention seems to have been dedicated to the composition of indexes and repertoires of bibliographic production through the ages. This does not mean there is little interest in KO. On the contrary, art history and historiography seem, in fact, to have always turned their interests towards sources that essentially show esthetic and formal principles, which means that objects such as paintings, sculptures and their reproductions have remained central to the study of art, being its main source and thus deserving all attention. The order and cataloguing of artworks and their reproductions are of central concern to bibliography, together with books, which are considered more relatively than in other fields of knowledge, where written books stand as the privileged, or even the only sources. In this paper, part of a wider research project, we will try to show how this KO practice originates from the era when the field designated “art” began to be developed according to a specific design. Authors involved in art criticism and theory within the
knowledge field of art history, such as Blunt (1966) and Venturi (1966), clearly show how the “architecture” of the art field was thought out during the Renaissance. It is during this time that we come across a discipline that can be defined as art history, clearly distinct from previous written records. We consider it important to reflect on these first statements on art, made during the Renaissance, in order to understand how and why the art field developed its own KO, which by the 18th century had achieved a systematic status (Strassoldo 2010): Art System created its own rules and established its own actors, and analyzing the sources of its foundation can help to better understand its mechanisms and its present directions.

The 16th century is considered the age that gave rise to this new field of knowledge called “art”, which was shown to need its own history. Historically, this is related to a major work by Giorgio Vasari, a mannerist painter, architect and the author of *Le vite de’ più eccellenti pittori, scultori, e architettori*, first published in 1550 (1986). This work is considered the first of its kind, dedicated exclusively to art and artist. There are, of course, numerous references in ancient literature to what we classify as art, such as paintings, statues or buildings. This way of writing about art essentially continued from antiquity into the Western Middle Ages, from which time survive accounts of the building activities of certain individuals, anecdotal descriptions of sites and manuals of techniques from the various arts, as extensively shown by Tosatti (2007).

The art texts we begin to recognize as art history writings in the Renaissance differ from these older kinds of writings in at least three ways:

1) In antiquity and the Middle Ages comments on art are all embedded in accounts of other subjects, whereas with the Renaissance, art becomes a subject in its own right.

2) There is an established tradition of writers on art history from the Renaissance to the present day, but not from antiquity.

3) One of the most formative ideas established in European culture, namely the cyclical image devised by Petrarch in the 16th century – whereby civilization reached a peak in antiquity, followed by a decline and then a revival, in the 14th century – was applied by Ghiberti to the chronology of art. According to Fernie (1996), Ghiberti states that Christianity had resulted in the destruction of all statues and pictures of nobility and perfection, such that art itself had ceased until Giotto’s talent revived it.

Vasari’s book *Le vite…*, first published in 1550 and in an expanded edition in 1568, is the first text on visual art and architecture extensive and consistent enough to be called a proper history. There is agreement, among scholars, that it was Vasari who set out the ground rules for history of art as a discipline, as it was followed for at least two centuries. The main elements of his approach can be grouped under the headings of “connoisseurship” and “humanism”, both acknowledging the central role of the idea of “genius”. Today, one cannot really speak of Vasari’s method as art history because he paid almost no attention to the historical and social context of what he was writing,
while stressing the achievements of each artist, frequently using anecdotal stories. Nevertheless, he opened a new direction, through studying documentary evidence and using techniques of connoisseurship.

Not until the 18th century was Vasari’s method of art history revised when Johann Joachim Winkelmann wrote a history of art that went further than merely being a history of artists.

On the premise that, as we stated, art history is based not only nor mainly on literature, but on the materiality (or representation) of art works, allowed by and rooted in the abilities of connoisseurship, we can observe that contemporary analyses in KO in the field of art (such as the Art & Architecture Thesaurus – AAT, a project linked to the Getty foundation that began in the late 1970s, or Iconoclass, a Dutch classification system planned as long ago as the 1950s) still tend to design, describe and organize art as a field in which art objects (their representation/reference) and art writings are not hierarchically distinguished. We think this deserves further research, and, in order to understand this particular scheme of KO in art history, we have focused attention on the way the field became explicit when it started to become autonomous, that is, when it emerged from the Renaissance.

Our inquiry revolves around the issues: how was the field’s conception also a result of written texts, of books? What about the constitution of bibliographic “maps” of authors and titles at the very beginning of the development of this field of knowledge?

We consider art a non-scientific field, related to practice and techniques of crafting, so it seems logical that authors who wrote on art favored this aspect of knowledge, as did Vasari. On the other hand, as we will see in the texts by Lomazzo and Possevino, another perspective began to appear in the field, more related to bibliographic discussion.

Without entering a theoretical discussion about bibliography, we would like to relate our paper to the idea of an enumerative bibliography, as Stokes (1982) defines the bibliography whose main goal is to create a document on all the existing material, or all that is known to have existed. From the perspective that bibliographies aim to provide scholars with research tools, we can focus on the author we choose to analyze in this paper. Our method consists in working on the original texts, so as to understand the choices in art literature that formed the basis and marked the frontiers of the field. Source studies combine with literary reflections to inquire what kind of knowledge field was designed by these authors and how it determined the development of art studies.

2. The author and his text: Lomazzo

Giovan Paolo Lomazzo (1538-1600) was a Lombard mannerist painter who moved from practice to theory when he became blind. One of the main results of Lomazzo’s reflection is the separation between art and art theory (Venturi 1966). In the writing we analyze here he presents a bibliographic review from antiquity until his age. The fourth
chapter of his *Idea del tempio della pittura* is a short enumerative bibliography containing a list of authors and titles of books, both in manuscript and printed form, but, above all, including examples of written texts as well as drawings and engravings. We believe that this demonstrates the central issue in a discussion that needs to be expanded with respect to KO: though art has always been considered an area in which unique pieces of visual art and architecture stand out, to a certain extent, it is a field in which there is a clear notion of the importance of documents that are not written. In this sense, art history, is the first discipline which, from its inception, has not been based on written texts alone.

We will now turn our attention to the text by Lomazzo.

Ancient writers: practically all the authors he lists are lost. They can be found cited by Pliny the Elder, but Lomazzo talks about them as if they could be found:

1) Menecmo. In Pliny’s Index: “Menaechmo qui de toreutice scripsit”. Toreutics is the art of working metals with hammer and chisel.

2) Lysippus. About him, Pliny again provides a reference, Durid of Samos (340-280 BC), disciple of Theophrastus – this is actually one of the references Pliny makes in his Index.

3) Pasitele. In this case, Lomazzo confuses this author with Praxitele, attributing 5 volumes on sculpture to him. Pliny writes: “qui mirabilia opera scripsit” on sculpture and among those “qui de pictura scripserunt”. And further: “qui et quinque voluminis scripsit nobilium operum in toto orbe”.

4) Euphranor of the Isthmus. Pliny cites him as an author on colours: “volumina quoque composuit de symmetria et coloribus”.

5) Antigonus. Pliny states that he wrote on toreutics and “qui volumina condidit sua arte”.

6) Xenocrates. Pliny says Xenocrat Zenonis wrote both on painting and sculpture, “Et de sua arte composuit volumina”.

7) Apelles. On this writer, Plinio says “voluminibus etiam editis, quae doctrinam eam continent” and “Apellis discipulus Perseus, ad quem de hac arte scripsit”.

After these ancient authors, whose works are lost, Lomazzo lists some references regarding mathematics and geometry:

1) Alhazen Ali Muhammad ibn al Hasan ibn al Haytham – (Bássora, 965 – Egypt 1039). Lomazzo explains that he who wrote 7 books, known since the 13th century thanks to a translation by a Polish monk, Vitellionis, the 3 books *De aspectibus* containing writings on light and the *Opticae Thesaurus*. Lomazzo probably knew them from the edition published by Risner in Basel, in 1572.

2) Vitruvius is, of course, cited by Lomazzo.

3) Caius Julius Iginus (60 a.C.-10 d.C.), who supposedly was a librarian during the empire of Augustus and wrote about optics.

4) Euclides. Lomazzo cites *I quindici libri degli elementi*. We can’t really know if
he refers to the edition translated by Cajano, published in Rome by Blado, 1545, or the complete works published under the title *Euclide*, translated by the famous mathematician Nicolau Tartaglia, in Venice, 1565.
6) Gemino. (a Greek astronomer).

At this point in his enumeration, Lomazzo states that nothing was written after these authors. In this sense he fully reflects the Renaissance vision of cycles of rise and decadence, subsequently going on to identify the modern authors on art. This part is particularly interesting, because he refers both to printed books and manuscripts as well as to sketch books, drawn and engraved:
1) Donato Bramante, a book of hand-made drawings, also cited by Vasari.
2) Bartolomeo Suardi (a.k.a. Bramantino), from Milan, who wrote various books on antiquities, not published, says Lomazzo.
3) Vincenzo Foppa, a Lombard painter, whose writings he saw, but were lost. This author is not cited by Vasari.
4) Baldassarre Petrucci (or Peruzzi). Lomazzo attributes him with authorship of a work on architecture, which, however, appears as a work by Sebastiano Serlio (*I cinque libri dell’architettura*).
5) Andrea Mantegna, a book containing drawings with explanations, a manuscript that Lomazzo saw in the collection of Andrea Gallarato, a noble abbot from Milan.
6) Bernardo Zenale, a treatise on perspective, published in 1524 and lost.
7) Bernardino Butinone, a collaborator of Zenale.
8) Marco da Siena, author of a volume on Architecture – also lost.
9) Leonardo da Vinci – a treatise on anatomy. Lomazzo states he saw it in the collection of Francesco Melzi, a pupil of Leonardo. (The master left Melzi all his books in a testament drafted in 1518). Lomazzo also refers to a part of the *Codex Urbinate*, in which the same Melzi reunited loose texts and thoughts of Leonardo, including some withdrawn from sources now lost. Lomazzo says none of these texts can be found printed, and that the manuscripts were given by Melzi’s son to the king of Spain’s sculptor, Pompeo Leoni. Later, they went to Guido Mazenta, brother of the architect and jurist Giovanni Ambrogii.

Lomazzo then offers a list of printed books by modern authors:
1) Leon Battista Alberti, *De Pictura* and, more relevant in Lomazzo view, the *De Re Aedificatoria*, 1450.
2) Pomponio Gaurico (who we also find in Possevino, but not presented by Vasari), *De sculptura*, Florence, 1504.
3) Luca da Borgo (a.k.a. Pacioli) *De divina proportione* – Veneza, 1497 and 1509.
4) Iacopo Barozzi (a.k.a.Vignola), *Regola dell’ordini d’architettura* (1562).
and *Le due Regole della prospettiva pratica* (posthumous, 1583).

Lomazzo further extends his references beyond Italian borders, listing foreign authors he considers significant:

1) Jan Steven van Calcar, author of the *planches* of Vesalius’ *De Humani corporis fabrica*.


**3. The author and his text: Possevino**

Jesuit Antonio Possevino (1533-1611), author of *Bibliotheca Selecta*, is considered “the most relevant catholic bibliographer of the 16th century” (Ceccarelli 1993, p. 713). His *Bibliotheca Selecta* offers an encyclopedic model rigidly delimited, that of catholic counter-reformed doctrine, it becomes an instrument of strict control of bibliographic information and of books circulation, aimed to build […] a collective memory selected according to a specific pedagogical program. (Balsamo 1984, p. 45).

Like Lomazzo, he proceeds chronologically in his narrative on art. Unlike the former, however, he doesn’t cite authors contained in Pliny as available sources. He only cites Pliny, using his volumes XXIV and XXV as the only source to provide a report on art techniques and artists. Possevino chooses to give bibliographic publications only when the sources can indeed be found. Such is the case of Philostratus of Lemnos, born in the 2nd century AD, author of the *Immagini*, or *Icone*, as Passevino explains. The work can be read in Greek, Latin and French, and the Italian version, by Giovanni Andreozzi from Lucca, had not yet been published. Our bibliographer explains that the Latin translation had been made a long time before, being published in Milan by Giovanni Castello in 1521. He also talks about the French translation, realized by Blaise de Vigenère.

A list of authors follows: Albrecht Düer and Pomponio Gaurico, also reported by Lomazzo; Pierre Gregoire, author of the *Syntaxes artis mirabilis*, an author put in the *Index librorum prohibitorum*; Giulio Cesare Scaligero, who wrote against Cardano (an author accused of being a magician).

From this point on, Possevino, who is not an artist, selects a set of authors and books that are not dedicated to techniques, but to morality in painting and sculpting. About each of these books he gives a short abstract:

1) Gregorio Comanini, a religious man from Lateran, author of *Il Figino*, published in 1591, in which Comanini discusses the aim of painting, whether it is useful or a matter of leisure and how to use it among Christians.

2) Giovan Battista Armenini, *De’ very precetti della pittura*, 1586, where issues on
how to paint places and people are discussed. Possevino highlights the preface, in which Armenini blames those artists who hide the rules of art so that without rules, learning is harder and slower. In his abstract, Possevino highlights moral qualities of art and artists in their duty.

3) Gabriele Paleotti, *Discorso intorno alle imagini sacre e profane*, Bologna, 1582. Possevino states that what the cardinal, “compelled by the ardent love for his diocese, wrote in his devote book about images is so important that no painter will ever regret to read it” (Possevino 1971, p. 53).

4) Andrea Gilio, who published a dialogue dedicated to cardinal Alessandro Farnese (1564), where he outlines the mistakes and abuses committed by painters in paintings stories. He also criticizes Michelangelo for his Final Judgement and other artists, explaining how sacred images must be.

5) Bartolomeo Ammannati, author of a letter to the *Accademici del disegno* (1582).

As we can see, Possevino offers a devotional review of the figurative texts of his century. It is interesting to point out that the last part of his bibliography on art is dedicated to the moral damage caused by the artistic representation of nudity.

4. Final remarks

The bibliographic repertoires by Lomazzo and Possevino are exemplar: both use ancient sources, opposing the criticism raised by Vasari, who was against their use (Pliny, Vitruvius etc.). Lomazzo and Possevino served to broaden the knowledge of authors working far from the Florentine circles, the former offering publications on Lombard books and authors, while the latter, more involved with moral concerns, highlights authors who do not necessarily “produce” art, but discuss it from a more theoretical point of view, being sensitive to the provisions of the Council of Trent (such as Gilio, Paleotti, and Comanini). Possevino does not consider Vasari among the authors.

It may be interesting to observe that Lomazzo, as a painter, is more concerned about sources (printed or not) that can be used by artists, such as works on geometry or perspective and sketch books. On the other hand, Possevino, a bibliographer, not an artist, worries about moral meanings expressed in art objects, suggesting books that depart from technique.

Considering that both authors are among the pioneers in compiling bibliographies on art, we consider them two important figures that can help us understand what kind of KO determines the field of art.
References

Gaius Plinius Secundus (s.d.). *Naturalis Historia*: liber I, XXIV and XXV. Available at: https://la.wikisource.org/wiki/Naturalis_Historia Visited the 29th of January of 2018.
Sonia Troitiño

Different parameters for Knowledge Organization in archives

Abstract
This article seeks to reflect on different knowledge organization parameters in respect of archives. It discusses the different ways we contextualize records and how retrieving this information can help with archival organization of documents, providing evidence of the source and preservation conditions undergone by an archive or collection throughout the trajectory of documents from production until their definitive keeping. In this regard, we believe data domain to be of utmost importance for institutional, administrative, archival history and custodial trajectory of records. As a result, record contextualization – one of archive science’s main purposes – based on distinct articulated perspectives, occurs more fully and integrally, offering the user full and consubstantiated access to the record.

The archival notion of knowledge organization is closely related to the process of functional contextualization of records. However, what is rarely discussed are the different parameters that can be adopted for knowledge organization (KO) in the sphere of archival science. Therefore, this paper proposes a reflection on the importance of recovering information regarding the history of the entity and the custodial trajectory of a file from creation of a document until it enters a permanent archive for the purpose of record, information contextualization and consequent knowledge organization relating to archival collections.

The archival notion of knowledge organization is closely related to the process of functional contextualization of records. However, what is rarely discussed are the different parameters that can be adopted for knowledge organization (KO) in the sphere of archival science. This paper proposes a reflection on the importance of recovering information relating to the history of the entity and the custodial trajectory of an archive from creation of a record until it enters a repository for processing, information contextualization and consequent knowledge organization.

Birger Hjørland (2016) states that

“KO is about describing, representing, filing and organizing documents and document representations as well as subjects and concepts both by humans and by computer programs. For these purposes, rules and standards are developed, including classification systems, lists of subject headings, thesauri and other forms of metadata. The organization of knowledge into classification systems and concept systems are core subjects in KO. The two main aspects of KO are (1) knowledge organization processes (KOP) and (2) knowledge organization systems (KOS). Knowledge organization processes (KOP) are, for example, the processes of cataloging, subject analysis, indexing and classification by humans or computers. Knowledge organization systems (KOS) are the selection of concepts with an indication of selected semantic relations”.

In addition, according to the author, the organization of archival knowledge should be considered part of knowledge organization itself (Hjørland 2016). On the other hand,
Mario Barité (2001) states that the first basic premise of knowledge organization is the understanding of knowledge as a product and social need. Dahlberg (1993) points out the important fact that knowledge does not exist without concept. Therefore, information retrieval from an information organization system depends on a pre-established logic of the concept itself.

When we consider different processes for organization of knowledge in the context of archival science, we are faced with more than one way to contextualize records. Therefore, choosing one kind of historical information as a parameter for archival processing, in the same way as recovering the knowledge of the systems for keeping custody of a set of documents that have the same origin, reveal the criteria adopted for archival organization.

Among the options considered here, we discuss the boundaries between institutional history/biography and administrative history of the entity of origin in the records as well as the archival history and custodial trajectory of records and how they lead to specific knowledge providing a greater understanding of the bureaucratic-social system of creation, use and filing of records. Dictionaries and glossaries of archival terminology do not usually define the above-mentioned expressions although we recurrently come across them in the specialized literature in the area. As a rule, the use of these terms, when used in texts on archives and documentation is more intuitive than based on methodological procedures of information and contextualization of records.

There are only a few reference works that address the topic. However, we sought to track some information for a definition. For this, we consulted Dicionário de Terminologia Arquivística (Camargo e Bellotto 2012), Glossary of Archival and Records Terminology (Pearce-Moses 2005), Lenguaje y Vocabulario Archivísticos (Heredia Herrera 2011), Diccionario de Archivistica (Cruz Mundet 2011) and others.

What we saw when we analyzed reference books is that very few of them define the terms to which we refer, though we find descriptions relating to them. In a way, this shows a gap in the bibliography, given that they are terms frequently used in the day-to-day collection of organization work, normally applied indiscriminately. Often, these expressions are incorporated into the methodology of archival work, such as for example in procedures that are essential for creating a classification scheme, in the context of national and international standards of archival descriptions, such as ISAD(G) (2000) and Nobrade (2006)

One of the sources found in the bibliography is referenced in the Diccionario de Archivistica organized by Cruz Mundet (2011). In the entry for history, the author mentions the existing relationship between history, archival science and their various facets. Overall, he establishes five points of intersection, mainly derived from those used by different description rules, which we reproduce below:

Administrative, institutional or biographical history: an element of ISAD(G) consistent with the synthesis of origin, evolution, development and work of entity or entities, person or physical
persons responsible for the production of document grouping is to put records into their context and ensure they are better understood, briefly annotating any significant information.

**Archival history**: an element of ISAD(G), which provides information on changes in ownership and custody of a descriptive unit that is significant. (Cruz Mundet 2011, p. 200-201)

As we can see in Cruz Mundet’s definition, the terms administrative, institutional, biographical and archival science were used together in two distinct definitions. In the first, despite their essential differences, concepts are equivalent when they assign the same function to their use. The second refers to the issue of ownership and custody; however, it does not mention interventions resulting from archival organization in its different phases.

It is important to note that, in both cases, the terms defined are shown as descriptive fields of ISAD(G), instead of being presented as elements to be retrieved within a mandatory methodological procedure for archival organization work. Thus, they normally refer to descriptive components – taking into account their position in the standard – rather than being understood as instrumental information that aids the archival knowledge organization process.

**Different histories, different perspectives for knowledge organization**

Recovery of institutional history is of fundamental importance for the organization of any archive since such history reveals bureaucratic practices and transformations encountered by entities throughout their existence. To get to know the entity organization system, identifying functions, activities and procedures, leads to the understanding of record system mechanisms and access to information and, consequently, to contextualization of record creation.

However, further to building a timeline, based on institutional milestones, other perspectives are involved and reveal themselves to be compelling in respect to the record organization process. We refer to a diversity of parameters adopted for recovery of an entity’s history and the custodial trajectory of its archive.

In archival science, institutional history is a very useful tool because it provides the knowledge of facts, elements and specific characteristics of a given administrative apparatus. It equally provides the knowledge of fundamental elements for creating an efficient information retrieval system as well as information on previous processes and phases in the archival organization of a set of documents.

According to Flórez (2010), institutional history analyses institutions’ organic and functional evolution and should be based on legal and chronological rules, taking into account the cycle in which the institution, which is the object of the study, is found, and also in the set of legal rules that structure it. Thus, according to the archival perspective, institutional history should encompass structure, function and evolution of organization, and is an integral part of the principle of origin.

What Flórez (2010) refers to as institutional history, in ISAD(G) appears as
“administrative/biographical history”, with a similar meaning and featuring as one of the elements present in the standard descriptive plan. According to ISAD(G), administrative/biographical history provides information that contributes to administrative or biographical history relating to the institution to facilitate understanding. As a rule, it guides the archivist to enter, in this descriptive field, any kind of information on the history of the collection to be described.

3.2.2 Administrative / Biographical history

Purpose: To provide an administrative history of, or biographical details on, the creator (or creators) of the unit of description to place the material in context and make it better understood.

Rules: Record concisely any significant information on the origin, progress, development and work of the organization (or organizations) or on the life and work of the individual (or individuals) responsible for the creation of the unit of description. If additional information is available in a published source, cite the source (ISAD(G) 2000, p. 18).

The administrative/biographical history field is one of the two intersection points between ISAD(G) and ISAAR(CPF), together with the name of the creator, who controls the authority registration. Both provide contextual information.

The elements of an Authority Record of ISAAR(CPF) – Identity Area, Description Area, Relationships Area and Control Area – suggest retrieval of specific information that may be included in the description of these elements such as full names and titles, dates of birth and death, place of birth, successive domiciles, activities, occupation or jobs, names of origin or any other, significant achievements and place of death for individuals or family.

This standard is intended to be used in conjunction with ISAD(G) - General International Standard Archival Description, 2nd edition and with national archival descriptive standards. When these standards are used together within the context of an archival descriptive system or network, authority records will be linked to descriptions of archives, and vice versa. [...] Descriptions of archives and records can be linked to archival authority records in the Name of creator(s) element (3.2.1) and the Administrative/Biographical history element (3.2.2) of an ISAD(G) compliant description (ISAAR(CPF) 2004, p. 12).

We do not intend to go deeper into the subject because this is not the theme proposed here; however, we have a reservation in respect of the many interpretations of the international ISAD(G) standard. Understanding what content should be addressed by descriptive elements, in the interpretations found for the standard, may lead to differences in the very act of description. For example, we mentioned that the Brazilian version of the standard does not refer to institutional history; instead, it refers to administrative history/biography. Whereas the Spanish version of the standard refers to the descriptive field as institutional history/bibliographic review; in English administrative/biographical history; and in French histoire administrative/notice biographique (ISAD(G), 2000). We can clearly see different approaches, according to the archival tradition of each country.
When we check the descriptive plans presented by various countries and incorporated into ISAD(G) to exemplify how a standard is applied to fonds of different origins and descriptive levels, the lack of uniformity in filling in information in relation to the history of the entity is clear. Standardization is more to do with the name of description than with the content to be described in these fields.

**Custodial trajectory faced with archival history**

On one hand, we have institutional and administrative history that provide us with contextualized information in respect to document production, routines, mission and social penetration of the entity – even serving as a factor of major influence to be considered at the time of document evaluation, especially when faced with primary and secondary values inherent to records. However, there is yet another kind of historical construction that must be considered at the time of contextualization of fonds: archival history.

When mentioning the difficulties in finding definitions for the various historical approaches for contextualizing a collection, we have cited the *Diccionário de Archivística*, by Cruz Mundet (2011), highlighting the fact that the approaches derive from archival description standards. Another reference work that also presents this question, at least partially, is the *Dicionário de Biblioteconomia e Arquivologia*, by Murílio Bastos da Cunha and Cordélia Robalinho de Oliveira Cavalcanti (2008). The authors define “administrative history, custodial history ADM ARQU, collection of information relating to history and structural organization of an institution” (Cunha and Cavalcanti 2008, p.186).

In this definition, we can see, especially through the equivalence of terms in Portuguese and English, a match between administrative history and custodial history. Normally, custodial history refers to the collection *per se*, whereas administrative history, in general, refers to the entity that processes documents. This position differs from that of Cruz Mundet (2011), who tends to treat the definitions of custodial trajectory and archival history as one.

Somewhat coincidentally, there is the guideline found in ISAD(G) for the term archival history, relating the term to information that is significant for authenticity, integrity and interpretation of the descriptive unit. As a rule, it registers information relating to origin and property, similarly to custodial history:

*Rules:* Record the successive transfers of ownership, responsibility and/or custody of the unit of description and indicate those actions, such as history of the arrangement, production of contemporary finding aids, re-use of the records for other purposes or software migrations, that have contributed to its present structure and arrangement. Give the dates of these actions, insofar as they can be ascertained. If the archival history is unknown, record that information (ISAD(G), 2000).

In this respect, archival history encompasses any type of information relating to
ownership, responsibility, custody and maintenance of a collection, as well as the information referring to documentary treatment used for its organization. As a rule, archival history depicts collection interventions relating to processes of organization, preservation and representation of records throughout its life cycle, including permanent interventions. In the cases where descriptive units have been acquired directly from the creator, the standard recommends that such information be registered as precedence rather than archival history because interventions suffered by documentation are the result of custodial procedures of the entity of origin.

Therefore, for ISAD(G), custodial history is an integral part of archival history. Accordingly, data on the custody of a collection should be addressed by archival history.

**Final considerations**

We saw that the distinction between administrative history and institutional history is based on the focus and perspective adopted as a reference. While the former is concerned with recovering organization and functional structures of a given entity and its reflection on document records arises from the execution of its activities, the latter focuses on a wider understanding of the entity, which includes identity, company mission and its role in society. Archival history in turn, is concerned with identifying and understanding the interventions relating to document organization made to a collection. Custodial history studies the trajectory of a collection from creation of records to their preservation and guardianship in close relation to an understanding of document ownership and custody.

The literature found and taken into consideration in this work, for the most part, corresponds to the definitions of history used in the phase of description for contextualization of records. A possible explanation would be perhaps a lack of discussion on the distinct types of contextualizing documents in other phases of archival processing. However, what we came to consider is that contextualization of documents and information is essential for organization of archival knowledge and can collaborate with all the other archival functions – Creation, Appraisal, Acquisition, Conservation, Classification, Description and Dissemination (Rousseau; Couture 1998).

As a result, we have seen throughout this survey that the choice of a parameter to be adopted for contextualization directly interferes in the completeness of the set of information to be used in archival organization and in search aids. Therefore, to speak of institutional history, administrative history or archival history means recovering distinct values that, when interrelated in a descriptive plan, have the ability to reflect more widely the context of a document within its life cycle than when used in isolation. This gives us indications to make us think that distinct dimensions of knowledge should be considered in the phase of information contextualization of an organization.
The survey results lead us to think that information relating to the history of an archive regardless of the parameter adopted reveals aspects that are fundamental for the processes of archival organization – especially those relating to classification and evidence of archival bond. Each one of these histories is equally capable of assisting users of fonds and collections to have access to documentation because it enables them to have access to more consistent and complete information about the collection they are analyzing.

References
Camargo, Ana Maria de Almeida e BELLOTO, Heloísa Liberalli (2012). *Diccionario de terminología arquivística*. 3ª ed. SP: ARQ-SP.
Marcilio de Brito, Widad Mustafa El Hadi, Maja Žumer, Simone Bastos Vieira, Marcos de Brito

Document representation with images: an experimental milestone

Abstract
Our previous research considered a document management system (DMS) interface from the perspective of an iOPAC (imaged OPAC) to meet the needs of a digitally adapted indexing process. The proposed indexing methodology makes use of a document/‘key-image’ correspondence in the user-system communication process, adopting the IFLA-LRM model. The present study intends to validate this conceptual model by establishing an experimental protocol. Using the same scope and methodology, the experiment was first applied to a group of deaf students and subsequently to a group of librarianship students, so as to contrast data and determine the interoperability variable. Confirming the working hypothesis, fifty-three per cent of the deaf population and eighty-eight per cent of the librarianship students recognised the relationship between a document and its ‘key-image’ while also considering it an easy and intuitive method.

1. Introduction
Images play a crucial role in knowledge organization systems and media applications in general (Sousanis 2015). Photography and cinema underwent an extraordinary development in the early twentieth century. At that time, Paul Otlet (1934, 224) discussed the sense of sight, whose field of exploration is broader than that of any other sense. Film, he says, acts on the brain even more directly than theatre, because it removes the effort of interpreting written and spoken language and concentrates emotion, by viewing things immediately. For Otlet, there has been a shortcut from reading or illiteracy, to cinema, since it is easier to see than to hear or even perceive language by its sounds.

An indexing model based on images requires reconsidering the current paradigm of using exclusively keywords to describe document contents. The ‘key-image’ concept establishes a legitimate approach to indexing documents with multiple perspectives in technical, professional and social areas. In this sense, document retrieval systems tend to adopt universal cognitive paths in knowledge organization systems (KOS) based on imaged communication, since they are no longer limited by the written language. This contribution to knowledge representation is supported by semiotic theories and presents an opportunity for documentation to evolve into the information technology (IT) society. The imaged indexing methodology can be implemented in user-focused systems compatible with the IFLA-LRM¹ model for social and cultural interoperability (Mustafa, 2015) throughout OPACs².

¹ IFLA – International Federation of Library Associations and Institutions; LRM - Library Reference Model (Salaba et al., 2011; Riva et al., 2015; Riva et al., 2017).
² Online Public Access Catalog.
Web search engines may use these alternatives for indexing techniques to support KOS performance. A broader area for interface development is now available with effective benefits for people with functional illiteracy such as deafness. In short, using images to create bonds between users and documents meet KOS challenges on a prosperous scientific basis.

Today the human-machine interfaces of document management systems (DMS) explore innovative solutions that may renew the concept of library, known as a cultural and social space which is close or distant, material or virtual. The envisaged solutions prioritise the organisations' mission rather than the technology employed. New libraries aim to be closer to their users by means of a set of characteristics identifying documental organisations attempting to “associate immaterial with material (for example the OPAC), local with distant (open/close/semi-open client-server architecture) and visible with invisible (semantic data on the bibliographical notes)” (Papy 2016, 13).

The concept of indexing with images outlined by Caribé and De Brito (2015) presupposes the use of intentionally composed images for document indexing purposes. Both indexing methods are complementary, and besides the role it plays in indexing languages, the ‘key-image’ based interface is becoming an imaged navigation tool which is far more intuitive than former written methods.

This study presents the results of an experimental validation step, as a milestone in the ongoing research project, demonstrating document/key-image correspondence while referring to information contents.

2. Presenting the study

Starting from the conceptual model of indexing with images (Caribé and De Brito 2015) the working hypothesis has been formulated to suggest that there is an “image-document” relationship from which the objects of the image refer to the objects of the discourse within a document, or a set of documents, in a relationship generally independent from the variable of orality, meaning independent from literacy. Considering that "visual literacy", which derives from gestures and images, is complementary to "oral-written literacy", they can be combined to minimise functional handicaps in communication.

2.1. The context

To verify the working hypothesis, an experimental context has been created in which the participants are, as far as possible, exempt from using their oral skills, meaning that they lack the speech or reading-writing mechanisms common to traditional DMS.

As developed by Marschark et al. (2000; 2004) knowledge organisation in deafness is a matter of mind. The oral communication variable is here interpreted as the physiological competences that lead to the development of oral and written languages. “Orality”, as designed by educational experts, is intrinsically linked to hearing because it is by depicting sounds that humans develop speech and consequently acquire the reading and writing competences proper to natural language. Sign language (SL), however, responds to those same properties without borrowing the same oral communication channel. SL is the visuospatial alternative for communication between
deaf people under a grammatical formal basis. SL users develop literacy which is appropriate for their linguistic universe and, therefore, sufficient for minimal integration in current social life. For these individuals, oral communication is deficient or total absent, and, moreover, incompatible with common social reading spaces.

We have thus defined an experimental protocol centred on the cognitive functions related to images and, at best, free from the interference of orality variables (reading-writing).

The existence of this cognitive image-document association is at the centre of our interest. To a certain degree these cognitive skills demonstrate the existence of intuition in the process. However, no particular concern has been emphasised about the nature of the process or the interpretative mechanisms used by individuals to achieve tasks.

Questions about reading access become clearer when viewed from the point of view of deaf education. Several scientific studies have dealt with this subject. Studies by Daniel Daigle (University of Montreal), Anne-Marie Parisot, and Colette Dubuisson’s (UQAM3) have led to a new understanding of how socializing with others is a motivating factor for attending school, as well as for making friends or developing psycho-social skills equivalent to those of hearing people (Dubuisson et al. 2009). Key to this social success, the most important factors are related to the development of language. Indeed, a language deficit is not always linked to deafness, but is due to some environmental variables that would cause a lack of accessibility to language. Luna et al. (2016) report on the dynamics of social power that plays an important role in interpersonal relationships and a certain sense of contempt for the hard of hearing, that destroys the fabric of trust, necessary for developing language skills, bilingualism and balance between diversity.

Among the influential oral communication factors, Dubuisson, quoted by Luna et al. (2016), highlights external factors such as: (i) the subject being treated; (ii) the competence of the interlocutor, (iii) the situation of the signer (e.g. hands occupied), and (iv) the communication intention. In addition, linguistic contacts with the family or the social milieu play a crucial role in the evolution of SL skills.

In her thesis about the bilingual and multimodal approach to orality in the deaf child, Isabelle Estève (2011) discusses other aspects that influence oral communication. According to her, a holistic and pictorial form of communication could fill expressive gaps, because the pictorial and linguistic organisation of thought forms part of an "integrated bimodal language system". These modalities represent the organisation of information that comes from a language that is preferentially gestural. In the childlike multimodality of the deaf, Estève recovers the links between vocal communication and gesture, evoking a vertical, dynamic and global component of statements that combine the linguistic and pictorial skills related to contents.

3 Université du Québec à Montréal (UQAM).
Therefore, from Estève's assertions (2011) it is inferable that there is an image iconicity for the lexicon as much in the sense of Saussure, a word sign, as in the sense of Peirce, a sign-image.

In addition, Perini and Righini-Leroy (2008) stress the importance of giving deaf learners access to writing practices. She draws our attention to the effectiveness of visual tools and multimedia support, for their capacity to present simultaneously text, images and statements, providing additional motivation for the learner. Illiteracy is basically the most common side effect of deafness, partly due to educational and school experiences that would produce an inappropriate blocking effect associated with some linguistic insecurity against the vernacular second language. To remedy this, the learner must gradually be brought to master reading and writing in a perspective of successful citizenship.

From a quantitative point of view, according to WHO\(^4\), more than 5% of the world's population, or 360 million people, suffer from disabling hearing impairment, \textit{i.e.} 328 million adults and 32 million children. Daigle, quoted by Tominska Conte (2013), reports that in the United States although there are 30,000 deaf students in higher schools, only a quarter of these students complete their studies. In Europe few deaf people reach High School level.

Through these studies it is formally recognised that the barriers of functional illiteracy are a handicap embedded within another handicap.

### 2.2. The experimental plan

The study population was divided into groups. The first group consisted of twenty-six individuals, all deaf, SL students, who volunteered to test our hypothesis. The control group consisted of thirty-one students in librarianship\(^5\).

The documental corpus included a batch of 50 scientific articles taken from the SciELO\(^6\) electronic library. The objective of this specialised corpus is to know whether the method of indexing with images applies to scientific documentation, since it is more formal, loaded in text and less illustrated. The documents in the SciELO library were obtained by simple queries on specific chosen domains, namely: Sports [football; handball; swimming; tennis; others] and Animals [crocodile, monkey, toucan; others]. The subclass "others" represents the elements of doubt, ambiguity or even out of context, often encountered in documental research in real situations.

Pre-processing the abstracts: We chose to summarise the texts making them easier for the interpretant in SL. Also, by simplifying abstracts, we intended to enhance text comprehension for a non-scientific public.

Composing the ‘key-images’: We applied the collage technique to a selection of

\(^4\) World Health Organisation. URL: http://www.who.int/mediacentre/factsheets/fs300/fr/
\(^5\) Faculty of Information Sciences, University of Brasilia-Brazil
images, from public domain (Google images) according to the pre-established procedure proposed by Caribé and De Brito (2015) and following Pierce’s (Nøth 2012) and Bertin’s (1970; Dantier 2008) semiotic principles.

Experimental proceedings: The abstracts were interpreted in SL and recorded separately in video format. They were randomly selected for the test. Each video corresponded to a document, both identified by their numbers. The ‘key-images’ were presented in PowerPoint format, organised according to classes (Sports and Animals). At the end, the participant’s choice was recorded on the answer form. The number of documents presented in the session was limited by the time available, which averaged around 100 minutes/session and 10 minutes/document, 16 documents in total for the deaf students. For the librarianship students, there was one session test with 19 documents in total, taking around 120 minutes/session and 6 minutes/document on average.

The answer form structure consisted of two parts: I) The individual’s profile: a) age; b) gender; c) level of education; d) degree of deafness; and e) competence in SL. II) The document-image association: a) choices in ordering class menus. Indirectly measuring the ability to follow hierarchical paths; b) Degree of personal difficulty with regard to the subjects and; c) Choice of 'key-images' best corresponding to the document presented (in LS).

2.3. Findings

The first group of 26 sign language (SL) students is heterogeneous in age, [18-25] 42%, [26-35] 23% and [36-45] 23%. They have a preponderance to being profoundly deaf (23% severe and 54% profound) and despite this significant handicap, the majority (69%) of students reported having high school level, and 27% having a university degree. Concerning sign language skills, the group is homogeneous at 92%, with strong SL experience.

The second group of 31 librarianship students is in the main represented by young people aged 18-25 years old (84%), studying at university (90%) or having a previous university degree (10%).

On a scale from 1 (very easy) to 5 (very difficult) (Erro! A origem da referência não foi encontrada.), the individuals’ perception about the subjects (scientific articles) was easy and moderately easy to understand (in SL) for 65% of them (level [1] 45% + level [2] 20%). On the other hand, for 75% of the library students the exercise was easy and moderately easy to understand (level [1] 54% + level [2] 21%). Corresponding to the rates of difficulty, exact matching of image and associated document was accomplished by 53% of SL students, followed by 88% of the library students.

In fact, the results of the exercise demonstrated that there is a correlation between these two entities, document and 'key-images'. The success rates of 53% of the SL students and 88% of the library students have confirmed our assumption about indexing
with images. The complexity of the document's subject and the cognitive mechanisms of retrieval seem to influence each other. This can explain how 39% of failure is related to the 27% (13% + 8% + 6%) of the overall difficulty for the SL students. Likewise, for the librarianship students, the 12% of failure is related to the 25% of perceived difficulty. The gap between these two variables may also have occurred due to external factors such as: a) inaccuracy of the indexing technique, b) image quality and/or choices relating to Google's image collection or c) misunderstandings due to the complexity of the subject. Regarding the effectiveness of the method, the average time spent on understanding the contents and finding and reaching a key image for each document was about 10 minutes for SL students and 6 minutes for librarianship students. Mastering the mechanisms of the retrieval system emphasises both the comprehension path leading to document content as well as the consistency of the document/key-image association process, thus testifying to the intuitiveness of the "indexing with image" tool.

Figure 1: Proportion of levels in perceived difficulty regarding document content, comparing SL and Library students

The proportionality graphs (Erro! A origem da referência não foi encontrada.) show the results by document for each group related to the document/key-image association achievements.

The unanswered questions indicate the presence of residual barriers, not solved during the sessions, probably due to the indexing method, to the subject's complexity, or else related to SL skills. This behaviour raises another question, namely the effectiveness of the message conveyed by the ‘key-images’ according to the linguistic competence of the interlocutors (Luna and Parisot 2016). It should be considered that this phenomenon, also found in systems with keywords, cannot be controlled downstream of the process without compromising the validity of experimental protocol. Overall, the results are significant, especially coupled with the enthusiastic appreciation from the students, recorded on video testimonies at the end of each session. This motivational variable meets the viewpoint of Dubuisson et al. (2009) concerning the...
appraisal of psycho-social skills motivating individuals towards reading practices.

Figure 2: Proportion of success in corresponding document with ‘key-image’, comparing SL and Library students

As documents were selected randomly, only a few of them happened to be analysed by both groups of students. In those cases, as can be seen in the graph (Figure 3), despite the SL students’ oral communication barriers, the results exceeded expectations, being very close to the contrasting group of librarianship students.

Figure 3: Comparison of text-image association with documents beyond groups

3. Concluding remarks

Our hypothesis revisits the issues of document and indexing concepts in a multidisciplinary and innovative approach. With a methodological basis in semiotics and discourse linguistics, it proposes adopting the indexing by image conceptual model
previously developed. It is now understandable that designing a KOS aiming to improve reading practices is a multidimensional research problem involving: document processing, data model, system interface and social and cultural interoperability. The experimental plan takes into consideration the theoretical contributions of the education domain regarding the study of reading by multimodal didactics concerning image (visual and oral communication).

By conducting the experiment with both hearing impaired and non-impaired student groups it was possible to evaluate variables such as the relevance of indexing with images, the importance of the visual in the document retrieval process and the difficulties facing orally challenged users in the information acquisition process.

The methodology emphasises the communication barriers caused by the absence of reading-writing skills. Moreover, since the SL group was not familiar with most of the scientific fields covered by the documents and as there was innovation in the method of knowledge organization, the experiment has brought up a new uncommon difficulty. Nevertheless, the results clearly demonstrated that the group was able to establish at first glance a univocal association between a scientific article and its corresponding ‘key-image’. This set of characteristics denoting efficiency and intuitiveness was pointed out by most of the students in favour of the envisaged information system.

Concerning interoperability, making a scientific document accessible to a heterogeneous, non-reader public requires breaking through an important functional barrier. It presents a new perspective on communication with library users in general, particularly aiding functional illiteracy. With this in mind, the next stage of study will be to apply this same experiment to a culturally and linguistically distinct public so as to endorse the interoperability paths explored herein.

Acknowledgements

We would like to express our gratitude to the Association of Parents and Friends of the Deaf (APADA) for providing workspaces and SL interpreters, essential resources for the conclusion of this experiment. We would also like to thank the SL students who so generously agreed to collaborate with this study.

References


Domain analysis of scientific production in Information and Communication Technology in the context of small farmers

Abstract
This paper presents an evaluation of the scientific production regarding the use of information and communication technologies (ICT) in the context of small farmers. Using domain analysis, it aims to define the structure by which such research is organized in the world. More specifically, it seeks to identify the main researchers in the area and the intensity of communication among them, through citation and co-citation studies. The methodology is characterized as a theoretical study through a literature review on the theme of the use of ICT in the context of small farmers, and as a bibliometric study conducted on the Elsevier SCOPUS database. The results obtained include description of the most cited authors and the identification of three clusters in the co-citation network.

1. Introduction
The new information and communication technologies (ICT), especially the Internet and the Web, are important alternatives for enabling the sharing of information content in an efficient and dialogical way. Viero and Silveira (2011) point out that the possibilities offered by ICT for the rural environment are diverse, highlighting their role in broadening horizons and in incorporating new expectations; establishment of marketing groups; new public policies; crop estimates and performance on the stock and commodity exchanges; bank services; credit and production cooperatives; as well as distance education and technical assistance. The rural environment is thus a diverse sector under continual development, increasingly demanding constantly updated information.

Some factors may hinder the use of ICT, such as geographic location and the current conditions of social and economic inequality in which small farmers are found. This can directly interfere with the use of technologies and access to information, preventing the informational needs of small farmers to be met due to the lack of informational content, consistent with their culture, the precariousness of the media and inadequate financial support for technology transfer in such rural communities (Meitei; Devi 2009). These factors make digital technologies more accessible to larger farmers, in a better financial position to acquire equipment and with greater skills in using access devices.

Recent Brazilian studies addressing the contributions of ICT to the small-scale agriculture sector include Viero and Silveira (2011), Miranda and Assis (2013), Sant'Ana (2017). These studies demonstrate the need to systematically situate research
involving the use of ICT in the context of small farmers, locating its core researchers and different strands of thought.

Considering the above, this study aims to define the structure by which such research is organized in the world, through domain analysis. More specifically, to identify the main researchers in the area and the intensity of communication among them, through citation and co-citation studies.

The relevance of this research lies in the privileged position Brazil occupies as the most productive country in terms of agricultural goods in Latin America, and eighth in the world ranking (SCImago Journal Rank 2017).

2. Domain analysis and bibliometric studies

Scientific production can be considered as input to analyze a particular domain. The concept of domain can be understood as an area of knowledge, a group of individuals working together in an organized way, or a set of publications (Mai 2005).

According to Hjørland (2002), a domain can be understood as reflecting a discursive community and its role in science. Its importance in the characterization and evaluation of science allows us to identify the conditions by which scientific knowledge is constructed and socialized.

For Hjørland (2002), a domain can be analyzed through 11 approaches, among them bibliometric studies, often associated with historical, epistemological, and critical studies, producing consistent results.

In this study, a bibliometric approach was used to analyze the scientific domain, whereby we researched the quantitative aspects of the production, dissemination and use of registered information, analyzed and contextualized in the group targeted for domain analysis.

In the field of bibliometrics, Gracio and Oliveira (2013) point out that the set of references from scientific papers can be analyzed as a reflection of a discursive community, as it reveals the authors recognized by this community, portraying characteristics of the domain. Their study is based on analysis of the frequencies of citations, whether of authors or documents, and of frequencies of co-occurrence – co-citations – among them. Such citation analysis thus seeks to visualize a scientific field as a domain, citations providing the main means through which the community of this domain legitimizes itself (Piovezan, Fujita 2015). This enables the relevant groups of scientists and their publications to be identified, and highlights the researchers with the greatest impact within an area (Glänzel 2003).

Associated with citation studies, co-citation analysis addresses the frequency with which two documents, authors, journals or countries, among others, are cited simultaneously in the scientific literature, based on the premise that when two documents or authors are cited together in a later paper, there is a closeness of subject
matter between those being cited, from the perspective of those who are citing them (Grácio; Oliveira 2013).

McCain (1990) points out that co-citation analysis allows the intellectual structure of a certain area or set of researchers to be studied, presenting the social and cognitive structure that binds them, that is, their domain of research. Miguel, Moya Anegon and Herreno Solana (2008) concur with McCain (1990) by highlighting that co-citation analysis, whether of documents, authors, journals or areas of knowledge, produces valid representations of the intellectual structure of a networked scientific domain.

A social structure in the form of a network is composed of nodes connecting members and channeling resources, focusing on the characteristics of the nodes to the detriment of the individual characteristics of the members, that is, as networks of individual relationships that people create, maintain and use in the course of their lives (Ote; Rousseau, 2002). These authors list indicators for social network analysis, including density, degree centrality, proximity, betweenness, among others.

3. Methodology

Our research is characterized as a theoretical study through a literature review on the theme of the use of ICT in the context of small farmers, and as a bibliometric study. The survey was conducted on the Elsevier SCOPUS database, on September 1, 2017. Two combinations of search terms were used, considering the words present in the title, abstract or keywords: [1] "small farmer" AND ICT and [2] "small producer" AND TIC. From this search, 15 articles were located, all through the first search, and were selected by reading the abstract, in order to confirm their relevance to the study.

The citation set of each article was downloaded, accessible through comma-separated values (CSV) files, whose references were consolidated into one single text file. The double citations of the texts were then removed, and citation counting was carried out to verify the most cited authors.

In total, the research sample consisted of 471 authors referenced in the 15 articles retrieved, of which 447 received only 1 citation (approximately 95%), indicating a large dispersion of cited authors. From the total number of cited authors, we selected those who obtained more than one citation to compose the sample of author co-citation analysis, totaling 24 researchers (approximately 5%). This proportion is close to the model of Price's Law, which proposes the square root of the number of cited authors (the root of 447 is approximately 22 cited authors).

For a qualitative evaluation of the results, we sought information about the 24 most cited authors to identify their nationality and institution, in order to provide a basis for the results obtained. From the matrix generated with the 15 articles by the 24 most cited authors, a 24 x 24 square matrix was generated with the absolute frequencies of co-citation among them.
The co-citation network was generated including the most cited authors using the software UCINET (v.6.636), and its NetDraw tool, generating a visualization that approximates the most similar authors in relation to the absolute values of co-citation. Finally, density and centrality indicators were calculated.

4. Results

The retrieval, restricted to 15 articles, with no specific time frame, showed that the issue of ICT and small farmers began to be observed with greater diligence by the academic community from the beginning of the year 2000. The distribution of publications per year, among the collected articles, are concentrated most highly in the year of 2014, followed by 2015, a fact that may be associated with a policy of the Food and Agricultural Organization (FAO/UN) which officially declared that year as the International Year of Family Farming. It is hypothesized that this policy may have motivated the increase in research on the subject.

Authors who received at least two citations in different articles (approximately 14% of papers) were selected, resulting in 24 authors. It is worth noting that 10 (66%) of the analyzed articles cited at least one of the 24 researchers, proving that they are representative in the field of ICT use in the context of small farmers, and, on average, each analyzed article cited 5.1 authors from this group.

Among the most cited researchers, the international literature clearly stands out, although Brazil is prominent in scientific production within the area of agriculture according to the SCImago Journal Rank. Of the 24 authors, all are of foreign origin, predominantly Asian (7 authors) and North American (4 authors). As for Europeans, there were 7 authors from Italy, Spain, the Netherlands and France (1 author) and two authors from the United Kingdom. There was also a presence of authors from Central America (3 authors), Africa (2 authors), and Oceania (1 author), with only South America having no authors among those most cited.

Regarding the institutions to which they belong, we identified three authors associated with the World Bank and three authors associated with the International Food Policy Research Institute (IFPRI), as well as authors linked to universities and other research institutes, among them Manchester, Yale and Magdalen College Oxford.

The three most cited authors, mentioned in at least three of the analyzed papers (20% of the total), are: Richard Heeks, a professor at the University of Manchester, who conducts research on the use of ICT for International Development (ICT4D); Keith Griffin, from Magdalen College, Oxford University, who has worked on the economic aspects of poverty reduction, with his books "The Political Economy of Agrarian Change: an Essay on the Green Revolution" and "Alternative Strategies for Economic Change: an Essay on the Green Revolution".

"Development"; and Ashok Jhunjhunwala, a professor in the Department of Electrical Engineering at the Indian Institute of Technology (Madras), a pioneer in promoting industry-academia interaction in India.

Table 1: Most cited researchers and their nationality and institution.

<table>
<thead>
<tr>
<th>AUTHORS</th>
<th>CITATION</th>
<th>NATIONALITY</th>
<th>INSTITUTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Griffin, Keith</td>
<td>3</td>
<td>Panama</td>
<td>Magdalen College</td>
</tr>
<tr>
<td>Heeks, Richard</td>
<td>3</td>
<td>UK</td>
<td>School of Environment, Education and Development.</td>
</tr>
<tr>
<td>Jhunjhunwala, Ashok</td>
<td>3</td>
<td>India</td>
<td>Department of Electrical Engineering</td>
</tr>
<tr>
<td>Abraham, Reuben</td>
<td>2</td>
<td>India</td>
<td>Indian School of Business</td>
</tr>
<tr>
<td>Asenso-Okyere, Kwadwo</td>
<td>2</td>
<td>Ethiopia</td>
<td>International Food Policy Research Institute</td>
</tr>
<tr>
<td>Babu S.</td>
<td>2</td>
<td>USA</td>
<td>International Food Policy Research Institute</td>
</tr>
<tr>
<td>Balit, Silvia</td>
<td>2</td>
<td>Italy</td>
<td>Food and Agricultural Organization</td>
</tr>
<tr>
<td>Bhatnagar, Subhash</td>
<td>2</td>
<td>India</td>
<td>Indian Institute of Management</td>
</tr>
<tr>
<td>De Freitas, Claudette</td>
<td>2</td>
<td>Trinidad and Tobago</td>
<td>Caribbean Agricultural Research and Development Institute</td>
</tr>
<tr>
<td>Evenson, Robert E.</td>
<td>2</td>
<td>USA</td>
<td>International Development Economics Program</td>
</tr>
<tr>
<td>Ganpat, Wayne G.</td>
<td>2</td>
<td>Trinidad and Tobago</td>
<td>University of West Indies</td>
</tr>
<tr>
<td>Glendenning, Claire J.</td>
<td>2</td>
<td>Ethiopia</td>
<td>International Food Policy Research Institute</td>
</tr>
<tr>
<td>Hall, Andrew</td>
<td>2</td>
<td>Australia</td>
<td>Food System Innovation, Partnering and Learning for Impact</td>
</tr>
<tr>
<td>James, Jeffrey</td>
<td>2</td>
<td>Netherlands</td>
<td>Center for Research in Economics and Business</td>
</tr>
<tr>
<td>Kenny, Charles</td>
<td>2</td>
<td>UK</td>
<td>Center for Global Development</td>
</tr>
<tr>
<td>Kirkman, Geoffrey</td>
<td>2</td>
<td>USA</td>
<td>Watson Institute for International Studies</td>
</tr>
<tr>
<td>Mittal, Surabhi</td>
<td>2</td>
<td>India</td>
<td>International Maize and Wheat Improvement Center</td>
</tr>
<tr>
<td>Navas-Sabater, Juan</td>
<td>2</td>
<td>Spain</td>
<td>World Bank</td>
</tr>
<tr>
<td>Pigato, Miria</td>
<td>2</td>
<td>Italy</td>
<td>World Bank</td>
</tr>
<tr>
<td>Qiang, Cao</td>
<td>2</td>
<td>China</td>
<td>College of Agriculture</td>
</tr>
<tr>
<td>Saravanan, Raj</td>
<td>2</td>
<td>India</td>
<td>National Institute of Agricultural Extension Management</td>
</tr>
<tr>
<td>Schweare, Robert</td>
<td>2</td>
<td>USA</td>
<td>World Bank</td>
</tr>
<tr>
<td>Shah, Tushar</td>
<td>2</td>
<td>France</td>
<td>International Water Management Institute</td>
</tr>
<tr>
<td>Yadav, K.P.S.</td>
<td>2</td>
<td>India</td>
<td>Maharashtra Industrial Training Institute</td>
</tr>
</tbody>
</table>

Source: By the authors.

Figure 1 shows the co-citation network of the 24 most cited authors, in which the size of the quadrilaterals reflects the intensity of degree centrality in the network, and
the thickness of the line segments is proportional to the frequency of co-citation between the authors it links. The image clearly reveals the formation of three clusters, denominated in this research as C1 (represented by the black quadrilaterals), C2 (represented by the blue quadrilaterals) and C3 (represented by the red quadrilaterals).

Regarding C1, it is noteworthy that it is composed of only two authors (De Freitas and Ganpat), both from Trinidad and Tobago, who have a paper in co-authorship in the book – *ICTs for Agricultural Extension: Global Experiments, Innovations and Experiences*. These authors were cited together with only two of the other authors, one of whom is Saravana, who edited this book, demonstrating a minimal degree centrality in the network of this cluster.

![Figure 1: Network of authors’ co-citation.](image)

Source: By the authors.

Regarding C2, composed of seven authors, we identify a strong relationship among three of them, who belong to the same institution, Asenso-Okyere, Glendenning and Babu, highlighted in the network by the intensity of the linking lines. There is also a strong proximity, in this cluster, among Indian authors Abraham, Mittal and Saravana, the North Americans Evenson and Babu, the Africans Asenso-Okyere and Glendenning, and the Australian Hall, indicating proximity of research on the subject among them.

As for C3, made up of 14 authors, the participation of Europeans Heeks, Pigato, Navas-Sabater, Shah, Balit, James and Kenny stand out, followed by Indian researchers Jhunhunwala, Yadav and Bhatnagar. In this cluster, the three most cited researchers
Jhunhunwala, Griffin and Heeks are present, two of whom, Jhunhunwala and Heeks, intersect the relations between C2 and C3, which demonstrates the relevance of these researchers for both groups.

It should be noted that C3 is composed mostly of authors cited in the two articles retrieved from Cecchini: "Can information and communications technology applications contribute to poverty reduction? Lessons from rural India" and "Information and communications technology for poverty reduction lessons from rural India", thus the strong relationship highlighted by the thickness of the lines in the network. These two articles were the oldest among the 15 analyzed, demonstrating a more classic strand represented by the cited authors, most of whom are European.

As for the density of the network, it presented a value of 0.819, or approximately 82% of cohesion among the researchers in this area. As for degree centrality, a 36% index of centrality was obtained, which demonstrates a network with a dispersed characteristic, a fact that may be attributed to the recentness of research on the subject, in which there is no consolidated bibliography. We highlight authors such as HEEKS (cited with 73.9% of the other authors) and Jhunhunwala (cited with 60.8% of the other authors), precisely the two authors that make the intersection between C2 and C3.

5. Conclusions

Our research sought to contribute to strengthening the understanding and use of citation analysis and author co-citation analysis as a procedure for domain analysis, evaluating the subject area of the use of ICT in the context of small farmers.

Three clusters were identified: C1 is composed of the proximity of Central American authors with little degree centrality; C2 of the proximity of Indian authors with those from Africa and North America; and C3 with a greater presence of European authors, followed by Indian authors.

We consider it important to expanding studies of this nature, addressing, for example, these same issues and using works retrieved in other databases, since the SCOPUS database did not retrieve any article in Portuguese. Although little consolidated, the recent growth of research interest in the use of ICT in the context of small rural properties is highlighted, represented by the increase of publications related to this topic from the last 15 years.
References
Epistemic communities, domain analysis, and Kuhn: dialogs and intersections in Knowledge Organization

Abstract
We study the concept of epistemic communities within the domain analytic approach in knowledge organization and we relate these concepts to Kuhn’s views on paradigms and other concepts of the sociology of science that are inherent to the foundations of library and information science.

Introduction
One of the core aspects of knowledge organization as a field is the organization of knowledge in classification systems and concept systems (Hjørland 2016). The social aspects of the production and consumption of knowledge are also inherent aspects to knowledge organization as these systems are a reflection of the social conditions and epistemology of its time, what Jesse Shera (1951, 82) called a “dependency of classification theory upon the state of the sociology of knowledge.” In the domain analytic approach to knowledge organization, which can be considered a solid continuation of the social epistemological approach established by Shera, disagreement is common and “the picture is really not one of agreement, but of conflicting schools, and the closer the neighbours the sharper the conflict” (Broadfield 1946, 69, cited in Hjørland 2016, 477). Within this approach, Richard Smiraglia (2012, 114) defined domain “as a unit of analysis for the construction of a KOS. That is, a domain is a group with an ontological base that reveals an underlying teleology, a set of common hypotheses, epistemological consensus on methodological approaches, and social semantics. [...] It is the interactions of the ontological, epistemological and sociological that define a domain and reveal its critical role in the evolution of knowledge.” This group has also been called an epistemic community in some studies in knowledge organization (Guimarães et al. 2015; Martinez-Ávila et al. 2017; Hjørland 2017).

Research question, hypothesis, objectives and methods
While domain analysis has been presented and used as a fruitful methodology in knowledge organization (e.g., Smiraglia 2015), our research question is: how can this approach be applied in the diachronic evaluation of knowledge organization systems and production of knowledge? Our hypothesis is that the theoretical framework of epistemic communities, and more specifically considering the dynamics of these communities, together with Kuhnian views on paradigms and scientific revolutions (1962), can assist the development and interpretation of diachronic studies in
knowledge organization. The objectives of this paper are: to develop the concept of epistemic communities within the domain analytic approach to knowledge organization; to discuss the dynamic aspects of epistemic communities in relation to domain analysis; and to relate these concepts to Kuhn’s views on paradigms and other concepts of the sociology of science that are inherent to the foundations of library and information science and knowledge organization.

We conducted a conceptual analysis of epistemic communities and other aspects related to our research question, such as domain analysis, domains, paradigms, scientific revolutions, and other concepts related to domain analysis and Kuhn in knowledge organization. We conducted a literature review of these aspects and discussed them in the context of studies that work with domain analysis as a methodological framework and diachronic studies in knowledge organization.

Results

Diachronic studies in knowledge organization include several possibilities, such as studies on scheme versioning of knowledge organization systems (see for instance the work on “methodological challenges in scheme versioning” by Joseph Tennis, 2016, and the special issue of the journal Knowledge Organization on “Subject Ontogeny and Knowledge Organization System Change”), and domain analyses working with bibliometric studies in various periods of time (e.g., Guimarães et al. 2015).

While production and interpretation of bibliometric studies is recognized as one of the main approaches to domain analysis (Hjørland 2002; 2017), Kuhn himself (1996, 177-178) also stated that communication networks can be studied (revealed) in the linkages among citations (and Kuhn here cited the works of Garfield, 1964, Kessler, 1965, and Price, 1965 as examples). Rolin (2017, 3) aptly dubs these networks as “specialty communities” (a term used by Kuhn elsewhere), and also quotes Kuhn to define “specialty communities” as “the producers and validators of scientific knowledge” (178). In this vein, the term specialty communities inevitably also evokes the domain analytic approach in information science, not only in relation to the definition of domain but also to other aspects such as the discussion on the relevance of subject-knowledge or information specialists (see Hjørland 2017). The domain analytic approach in Kuhn also seems to be present in the following sentence: “If, as I have already urged, there can be no scientifically or empirically neutral system of language or concepts, then the proposed construction of alternate tests and theories must proceed from within one or another paradigm-based tradition” (Kuhn 1996, 146).

Epistemic community is a concept introduced by German sociologist Burkart Holzner (1968). In the words of Håkanson (2010), “Epistemic communities consist of individuals with identical or similar ‘frames of reference’ and cognitive ‘orientation systems’. These are associated with specific social roles, such as those of different occupational groups, and are acquired in a process of cognitive socialization, usually
through a combination of formal training and on-the-job experience” (p. 1.807). Epistemic communities have been applied to a wide range of areas, from political science (Haas 1992) to sociology (Meyer and Molyneux-Hodgson 2010), among others. Its first use in information science, to the best of our knowledge, was in Guimarães et al. (2015). Epistemic communities can be studied for a given domain as they can be understood as networks of knowledge-based experts that “not only hold in common a set of principled and causal beliefs but also have shared notions of validity and a shared policy enterprise. Their authoritative claim to policy-relevant knowledge in a particular domain is based on their recognized expertise within that domain” (Haas 1992, 16). In philosophy, the link between these scientific communities, as described by Kuhn, and bibliometric studies has been termed “epistemic interest communities” (Massey 2014).

The correspondence between epistemic communities and science communities is not always obvious. Although the concept of epistemic communities has been related to scientific communities (e.g., Meyer and Molyneux-Hodgson, 2010, who state that “Scientific communities matter. The enduring importance of journals, conferences, university departments, peers, discussion lists, museums, as well as the formation of new kinds of collectives around new problems and mediums prove this point”), it has also been often applied to the study of business firms and other kinds of organizations (e.g., Håkanson, 2010) mixed with concepts from knowledge management such as “tacit knowledge” and “explicit knowledge” (Nonaka 1994). If the concept of epistemic communities has until now mainly been associated to this context, one possible argument to make it relevant for our discussion on domains and knowledge organization would be the interpretation of tacit knowledge by Thomas Kuhn (based on Michael Polanyi, 1958). Accordingly, it is said that much of the scientist’s success would be dependent upon knowledge that is acquired through practice and not articulated explicitly (albeit guided by a paradigm). Likewise, several authors have highlighted the resemblance between the concepts of “epistemic communities” and “communities of practice” (e.g., Håkanson, 2010, Meyer and Molyneux-Hodgson, 2010). In this vein, based on the work of Holzner (1968, 51-59, Holzner and Marx, 1979: 103–106), Håkanson (2010, 1807) also pointed out that the reality constructs accepted by an epistemic community reflect the epistemic criteria of validity and reliability that it subscribes to, recognizing the empirical tests and methodological procedures employed in scientific communities as one example.

The concept of epistemic community has been likened to Kuhn’s paradigm (Meyer and Molyneux-Hodgson, 2010) and also to scientific community (Cross 2013). The relation between paradigms and scientific communities in Kuhn is not trivial. In the 1969 postscript, Kuhn clarified that “a paradigm is what the members of a scientific community share, and, conversely, a scientific community consists of men who share a paradigm […] the practitioners of a scientific specialty” (Kuhn 1996, 177). In the first edition of the book, Kuhn had stated that paradigms are “universally recognized
scientific achievements that for a time provide model problems and solutions to a community of practitioners” (1996, x).

In the context of epistemic communities, Meyer and Molyneux-Hodgson (2010) summarize the relationship between paradigms and scientific communities as follows: “Kuhn's definition of science assumes a clear demarcation around and within the scientific community. For Kuhn to be a member of a scientific community, to be accepted as a scientist, is to accept and to work within a paradigm. The study of these paradigms is what 'mainly prepares the student for membership in the particular scientific community”.

We can also link Kuhn’s concept of paradigm to the bibliometric approach to domain analysis if we consider the following definition of paradigm by Kuhn as achievements that share the following two characteristics: “Their achievement was sufficiently unprecedented to attract an enduring group of adherents away from competing modes of scientific activity. Simultaneously, it was sufficiently open-ended to leave all sorts of problems for the redefined group of practitioners to resolve.” Kuhn also added that these achievements “are recounted, though seldom in their original form, by science textbooks, elementary and advanced. These textbooks expound the body of accepted theory, illustrate many or all of its successful applications, and compare these applications with exemplary observations and experiments” (Kuhn 1996, 10). The importance and problems of textbooks is discussed by Kuhn throughout his work and is linked to the concept of “normal science” (a moment prior to a scientific revolution). In fact, Kuhn seems to talk about a rudimentary version of a domain analysis following a bibliometric approach when he talks about the determination of a paradigm in the following terms: “Close historical investigation of a given specialty at a given time discloses a set of recurrent and quasi-standard illustrations of various theories in their conceptual, observational, and instrumental applications. These are the community's paradigms, revealed in its textbooks, lectures, and laboratory exercises. By studying them and by practicing with them, the members of the corresponding community learn their trade. The historian, of course, will discover in addition a penumbral area occupied by achievements whose status is still in doubt, but the core of solved problems and techniques will usually be clear. Despite occasional ambiguities, the paradigms of a mature scientific community can be determined with relative ease” (p.43). In this vein, the analysis and characterization of any textbooks and literature before or after a scientific revolution could also be used to define the boundaries of the epistemic community at a given moment, while the transformation of the paradigm (and the literature) with each revolution can reveal the dynamic transformation of the domain.

The transformation of the group (arguably an epistemic community) is discussed by Kuhn in the following passage: “it is sometimes just its reception of a paradigm that transforms a group previously interested merely in the study of nature into a profession or, at least, a discipline. In the sciences (though not in fields like medicine, technology,
and law, of which the principal *raison d'être* is an external social need), the formation of specialized journals, the foundation of specialists' societies, and the claim for a special place in the curriculum have usually been associated with a group's first reception of a single paradigm” (p. 19). In this regard, for instance, the foundation of societies such the ISKO or the journal “Knowledge Organization” could be seen as indicators of the emergence of a domain, in this case the domain of knowledge organization.

On the other hand, although it is true that Kuhn’s work mainly focused on hard science (understandable given that he was a physicist), we believe many of the concepts and ideas he posited can be applied to other domains and disciplines. Therefore, if we accept the existence of other fields and domains than the natural sciences, Kuhn seems to leave the door open to define them even if only by exclusion (“We have already seen, however, that one of the things a scientific community acquires with a paradigm is a criterion for choosing problems that, while the paradigm is taken for granted, can be assumed to have solutions. To a great extent these are the only problems that the community will admit as scientific or encourage its members to undertake. Other problems, including many that had previously been standard, are rejected as metaphysical, as the concern of another discipline” 37). It should be remembered that, even if Kuhn did not feel comfortable with the application of his ideas to the Social Sciences and Humanities, he did not hessite to cite Wittgenstein, Bacon, and other social scientists to support his arguments.

**Conclusion**

We believe that our hypothesis seems to be confirmed since epistemic communities can be studied in a dynamic way in order, for instance, to evaluate the development of knowledge organization systems or the transformation of domains in time. In this sense, Kuhn’s theoretical framework seems to be a valuable methodological resource for the interpretation of results. As an example, Guimarães *et al.* (2015) divided the domain analysis of the epistemic communities in knowledge organization into periods of five years, revealing a shift from an epistemic community led by Ingetraut Dahlberg (some might say theoretically driven by ontology) to an epistemic community led by Birger Hjørland (in which the epistemological/domain analytical approach seems to prevail). Using Kuhn’s terms for interpretation of the results, it could be said that there has been a paradigm shift within the knowledge organization domain/epistemic community; that is, a scientific revolution, after a period of competition in which a previously accepted theory was rejected in favor of a new one.

**Acknowledgment**

References


Tatiana de Almeida, Gustavo Silva Saldanha

Epistemic LOCI: linguistic and critical meta-methodology in Knowledge Organization

Abstract
The present research seeks to apply a cultural approach in the epistemological and historical construction of representational language practices. We have conducted a remote meta-methodological study of KO, specifically, by adopting a meta-methodological orientation plan from Emanuele Tesauro (1670), which highlights the importance of production in KO. This approach, herein understood as epistemic loci, seeks to reposition the concept in a specific discursive dynamic, revealing its actions to be the result of collective social data, in a given space, at a given time. The present study aims, overall, to understand the environments within which concepts are formed, consolidated, dispersed and appropriated. Thus, the purpose of the study focuses on the relationship between KO methodology and the construction of thought in KO, from a reflexive meta-methodological point of view. From Tesauro (1670) through to García Gutiérrez (2011), the basis for constructing epistemic loci has been framed within a philosophy of language, central to the development of a critical and discursive approach to KO. We can observe how discursive approaches to KO have an epistemological and historical root linked to the study of loci. In the same way, we can say that political and social questions, as well as the critical approach to sciences and their classification (an epistemological critique), find in the theory of commonplaces a clear power of argumentation and historical revision. In other words, using the notion of loci, we hereby construct another meta-methodological category for reflection on science, and on the epistemology of KO itself, namely epistemic loci.

1. Introduction: from loci to epistemic loci
How is it possible to recognize the role of culture in the epistemological-historical foundation of knowledge organization (KO)? Can we understand the historical presence of cultural elements in remote sources of KO construction? What methodologies in KO can be used to this aim? And what methodologies have been developed objectively for such a cultural approach?

These concerns arise in the light of key issues regarding epistemology of KO, as summarized below:

a) The contemporary emergence of social and political debate regarding KO practices demonstrates the need for approaches, methods and concepts to understand the cultural dynamics of society (Guimarães, Pinheiro and Milani 2006);

b) The construction of social and critical approaches to social problems, such as the approach to declassification by Antonio García Gutiérrez (2011, 2003).

c) The revision of historical epistemological approaches that contributed to this foundation, as in the case of the review of Ranganathan's work (Saldanha 2014).

The present research seeks to apply a cultural approach in the epistemological and historical construction of representational language practices. We have undertaken a
remote meta-methodological analysis of KO. Specifically, we have discovered in Emanuele Tesauro (1670) a meta-methodological orientation plan, which highlights the importance of production in KO.

In Tesauro (1670) we find a dynamic intersection of approaches merging the lines of analytical and discursive comprehension (a reading of Aristotle’s Organon, and Aristotle's Rhetoric / Poetic), including loci, a rhetorical method developed by Aristotle concerning common objects of ethics, epistemology and other branches of knowledge.

The loci approach, used this way, is not new in Information Science or in KO. It can be found, for example, in the classification of knowledge in Conrad Gesner (Araújo 2017), in baroque thinking in the organization of knowledge (Almeida and Crippa 2009) and in contemporary reflection on KO itself, as seen in García Gutiérrez (2011).

In this sense, the purpose of the study is based on the relationship between KO methodology and construction of thought in KO, from a reflexive meta-methodological point of view.

2. Epistemic loci as a meta-methodological approach to KO

The approach here, understood as epistemic loci, seeks to reposition the concept in a specific discursive dynamic, which reveals its actions as the result of collective social data, in a given space, at a given time. It aims, overall, to understand the environments within which concepts are formed, consolidated, dispersed and appropriated.

The notion of epistemic loci (the conceptual basis underlying the context of the present study) enables an approach to KO that supports the relationship between an analytical and a socio-discursive approach to the object. The strategic notion of place is an idea worth discussing here, as it takes the concept of space beyond its immediate scope, seeing it as extension or even duration.

Epistemic places depart, in rhetorical terms, from the demarcation of “tropes”, forms of fixation of language and its configuration in verbal signs. According to the Foucauldian definition, rhetoric is the primordial function of language, in dialogue with grammar:

[...] Rhetoric, which deals with figures and tropes, that is, how language is spatialized in verbal signs; [...] Rhetoric defines the spatiality of representation, as it is born with language [...] (Foucault, 2002, p. 116).

The notion of "space(s)" or "place(s)" has a constraint on the structure of language: it is a question of identifying in space-time the modes of fixation in a given discourse, concept, approach, method – or simply, in Foucault’s lexicon, the "spatiality of language". The present study therefore aims to study the tropes that demarcate KO today, an approach already indicated by Tesauro (1670).

Commonplaces or tropes thus highlight the essential "primitiveness" of a concept, as well as its experience in the contingent actuality that underlies its delicate organic
structure, its spatio-temporality always being at risk. Here, again, notions of space-time do not have absolute equivalence with questions of extension-duration. Rather, the theoretical plan of such a discursive cultural approach potentially enables one to understand the sociocultural environments of the manifestation of a given concept and its modes of repercussion and appropriation.

In this way, the processes of contextualization that mark discursive approaches are not merely based on language itself. Discourses are positioned, that is, situated in a space-time historicized by a set of social determinations.

3. The construction of social and critical approaches to social problems in KO practices

In the context of KO, various studies have demonstrated the ideological condition of documentary language construction. This is the case of the academic studies by Olson (1999, 2001), which address diversity and cultural issues in KO; Pinho (2010), which examines the representation of male homosexuality in documentary instruments, and Simões (2010), which discusses modes of understanding the development of the notion of "ethnicity" and its unfolding in categories with respect to Universal Decimal Classification (UDC). Critically, we can identify a political positioning arising from the construction of a social memory legitimized by an aggregate of discourses that represent authority.

In this case, the potential of information retrieval, that is, the relation between precision and recall, is met by the constant increase of inequality manifested in the impossibility of identifying non-hegemonic contents in databases.

A fundamental theoretical resource in this dialectical panorama of information retrieval is found in the "declassification" discourse of García Gutiérrez (2011), which seeks to understand an epistemology capable of overcoming the fractures of inequality manifested and multiplied by documentary languages. The Spanish author indicates metonymics as a tool that identifies a part within a set.

In his critique of the Western view of classification, which focuses directly on how we structure theoretical and applied elements in KO in order to optimize information retrieval, García Gutiérrez (2011) demonstrates how the metonymic resource is used as a political tool of forgetfulness. It is a principle of metonymic reduction that simplifies and fragments the plurality of cultural expressions, dissolving the processes of comprehension in their amplitude and favoring the dominant view of a given established rationality.

As the Spanish researcher has shown, we are generally guided by a starting point that is wanted and said to be non-ideological or neutral, supported by a technical-scientific discourse separating truth and culture. The proposal of a declassification point of view is, according to García Gutiérrez (2011), the inclusion of the pluralism of
knowledge and expressions in the logic of the organization of knowledge. In his words,

It is a metacognitive and non-automatic operation that, in each action of the classifier, requires a complete awareness of incompleteness, bias, and explicitable subjectivity. With current technology, it is possible to elaborate procedures and systems of classification based on declassification. But such techniques and tools will also have to undergo an epistemological revolution in all their protocols and strata (GARCÍA GUTIÉRREZ 2011, p.10).

The act of "declassifying" is not, contrary to what a preliminary view may conclude, the denial of "classification" or the nihilistic conditioning of information retrieval, which posits the absolute inaccessibility of data in any and every informational search. No: "declassifying" is only, García Gutierrez (2011) explains, the anti-dogmatic and partial affirmation of the dialectical (and not contradictory) relationship of the practices of knowledge organization

4. From methodological practices to meta-methodological study in KO

It is well known that KO itself is composed of several processes and activities that provide the means to analyze, organize, represent or classify any domain of knowledge, theme or mission-problem. In this way, we will appropriate the concept of meta-methodology in order to accomplish this task.

The recent meta-methodological practice necessarily relates already existing techniques, sophisticating and expanding its applications. The metamethodologies challenge uniform traditional practices, question some standard procedures to objectively present new experiments. There are no limits on the interactions, only those of the capacity of realization of its researchers (Miranda 2003, p. 162).

Based on the previous propositions, we will appropriate the concept of meta-methodology in order to analyze KO through its own processes and theories.

From Emanuele Tesauro (1670), the epistemic loci put forward discursive questions such as who speaks, where they speak, when they speak, to whom they speak (all of them being treated here as "places" where "commonplaces" are consolidated in language, or discursive spaces).

The dynamics of epistemic loci demonstrate that the construction of a "place" is based on a network of conditions of discourse, which are empirically manifested by characteristics such as institution, ownership, gender, social class, financial capital and theoretical trajectory. The place of a concept is only the apparent spatial figuration of a constant movement, our (in)ability to see “dance in immobility”. In the social and interpretative context of cultural reality, words and things of the epistemic loci, that is, of social, historical and contingent dimensions that (temporarily) determine the affirmation of one and the negation of another concept, the similarity or identity between the terms, the struggle for a "Space", for a "highlight in the landscape". The epistemic loci are the opening for the pre and post-analytical understanding of the concept, without denying its conceptual intentionality.
These conceptual demands and their approaches have been placed within the long tradition of methodological demarcation from so-called "discourse analysis" – as discussed by Olson (1999). Before discussing the effects of discursiveness from recent theories of discourse, such as that of Michel Pêcheux, the present study seeks to shift presuppositions of rhetoric itself onto another metamethodological plane, following the lead of Tesauro (1670).

5 Final remarks: a linguistic and critical epistemic metamethodology

The premises for the construction of epistemic loci are structured in a philosophy of language, central to the development of a critical and discursive approach of KO, from Tesauro (1670) to García Gutiérrez (2011).

As far as discourse is concerned, we recognize the existence of a contemporary body of research, involved with the approximation between discourse analysis and KO theories. In general, the notion of social indexation tends to embrace such approaches. In our case, it is about recovering the rhetorical thinking of tropes and rethinking the appropriation of the notion of commonplaces in a broad critical-discursive plane. We are, however, attentive to the epistemological-historical point of view of Emanuele Tesauro, already influenced by the Aristotelian perspective, as well as to current epistemological approaches, such as those envisioned by Rafael Capurro, García Gutiérrez and Holpe Olson.

Our research has followed two paths of thought until the point where they intersect. The convergence of the two metametodological dimensions will lead to a critical-historical understanding of the national development context of KO.

We can see how discursive approaches to KO have an epistemological and historical root linked to the study of loci. In the same way, we can say that political and social questions, as well as the critical approach to sciences and their classification (an epistemological critique), find in the theory of commonplaces a clear power of argumentation and historical revision. In other words, from the notion of loci, we construct here another metametodological category for reflection on science, and on the epistemology of KO itself, namely epistemic loci.

Acknowledgment

This research was developed under the financial support of the National Council for Scientific and Technological Development (CNPq), Brazil, and the Carlos Chagas Foundation for Research Support from Rio de Janeiro State (Faperj), Rio de Janeiro, Brazil.
References


Rosa San Segundo, Maria Adelina Codina-Canet

Epistemological challenges in Knowledge Organization in the digital age

Epistemology emerged in European modernity as a construction of symbolic order, establishing itself as a way to legitimize knowledge and as a strategic means of controlling it. Scientific knowledge has gone through different stages and has not always been built on the same epistemes, having been subject to constant epistemological changes. It is necessary, therefore, to analyze the experience of order from the experiences that exist in every culture. Our values and the way we think and know are integrated in the framework of the episteme in force, in the age in which we live. Discursive practices may seem free, but they are strongly conditioned by the epistemic structures that embrace them.

Technological postmodernism has emerged, based on a fundamental switch to the primacy of technology over science, and specifically of digital information technologies, with a new virtual world of electronic information, enabling the transmission of an overall vision of both cultural and historical experience. Thus, since the beginning of the 21st century, digital information technologies have become the driving force of the information and knowledge society, an agent of socio-economic change and a scientific discipline par excellence which has led to the technological determinism current in today’s world.

Accordingly, a global interconnection of information and semantic information is generated, which can also produce an exponential pattern of technological growth. In our postmodernity, the digital network has triggered revolutions in a variety of contexts, whether mediatic, scientific, epistemic, symbolic or organizational.

Objectives

Epistemology is the central discipline in the construction of scientific knowledge, addressing the fundamentals, criteria, and validation through which knowledge is justified. It is a metadiscourse dedicated to order and scientific analysis. Epistemology emerged in European modernity, as a construction of symbolic order, establishing itself as a way to legitimize knowledge and as a strategic means of controlling it.

In the history of science and culture, in every age, there is always an underlying order that regulates knowledge. This silent order is what enables and adjusts the constitution of knowledge – that knowledge contained in each moment. Thus, scientific knowledge has gone through different stages and has not always been built on the same epistemes, having been subject to constant epistemological changes. It is necessary, therefore, to analyze the experience of order from the experiences that exist in every culture. Our values and the way we think and know are integrated in the framework of the episteme in force, in the age in which we live. Discursive practices may seem free, but they are strongly conditioned by the epistemic structures that embrace them.

This paper thus aims to illuminate the configurations that have given rise to different forms of knowledge and ways in which scientific knowledge is produced in each epoch, keeping in mind that the fleetingness of epistemology and current volatility necessarily
leads to the dissection of the old epistemology.

Moreover, all cultures are immersed in and constituted by the differing technical levels that fostered and shaped them (García Gutiérrez, 2002). The scientific perspectives imposed in the age of science are very different. Currently, in the context of epistemology, the paradigms of chronologically replaced processes have given way to new paradigms which have, in turn, led to paradigms of the paradigms, such that the processes of scientific methods obey a universal process governed by an interrelation of generative factors.

Thus, all epistemological changes and evolutions do not obey a linear sequence, but rather are factors of a generative process. Likewise, knowledge is related to discursive communities, with the different social groups that make up modern society. The digital network allows for the opening of more non-exclusive ways of identification, so that culture and access to it has been democratized. As a result, the re-expropriation of symbolic and socially relevant space becomes necessary. The process of social belonging will be further opened with the creation of new communities constituted digitally from different groups and associations of very diverse nature. For this reason, dynamic models of organization must be proposed, avoiding exclusion and invisibility, towards integration and transculturality.

It will be necessary to incorporate a new perspective to be able to appreciate the magnitude of this mutation towards a new electronic episteme, which also points to a post-memory condition where meaning will be linked to the assembly of discontinuous and mobile elements. Meaning will be found in the relationship to the new supporting foundation that transmutes everything.

**Method**

At the beginning of the 20th century, Western European culture and knowledge, with its scientific models and science-based truths, pointed to an epistemology based on objectivity, which became integrated into the new status quo. The transition to postmodernity occurred at the end of World War II, through the creation of the digital technological world, bringing about a shift in that objectivity, rationality and cognoscibility, where nothing is alien to its process or time and, therefore, nothing is totally objectifiable.

Likewise, technological postmodernism has emerged, based on a fundamental switch to the primacy of technology over science, and specifically of digital information technologies, with a new virtual world of electronic information, enabling the transmission of an overall vision of both cultural and historical experience. Thus, since the beginning of the 21st century, digital information technologies have become the driving force of the information and knowledge society, an agent of socio-economic change, a scientific discipline par excellence, which has led to the technological
determinism current in today’s world.

In addition, the digital interaction characterized by the development of technologies aimed at participation and collaboration between virtual communities, with social networks that make up communities wherein users interact, will, in turn, lead to the establishment of a digital identity space. In this way, a global interconnection of information and semantic information will be generated, which may also produce an exponential pattern of technological growth.

A participatory culture has emerged with blogs, wikis, and social networks, where individual creativity operates. There is already a new way of thinking in the digital era that encompasses hybridization of materials, formats and texts. Just as forms of primitive thought influenced knowledge organization, writing also changed it, in the same way that digital thought is now shaping our reality and its organizational form. The web demands new ways of learning and knowing, such that digital information, connectivity, virtuality and hypertextuality are already part of human thought.

We are in the preeminence of the visual, the digital and now the virtual, our most visual environment, which is becoming increasingly more widely implemented. Virtual visualization systems, which provide the illusion of immersion within an image, is a rapidly evolving field, which is modifying our whole reality, and creating new philosophical and ethical problems caused by the progress of virtual imagery, as it generates consequences for our new way of representing and interpreting the world.

The globalization of information, connectivity, virtuality and hypertextuality are already part of human thought. The natural memory considered as a process that requires the structure of natural language and human capacities, is interwoven in subject and consciousness. Memory saved through the medium of writing required the assistance of other tools, while today the multiple forms adopted by memory are invariably digital. The new tool used to recover digital memory is the semantic web, which points to the future of information on the Internet and seems to approach the utopia of organized global information, in the attempt to give more meaning to the web. Web 2 requires new ways of learning and participating. Web 3.0 incorporates the proliferation of languages, concepts and hard-to-manage tools created by users. The semantic web seems a natural evolution of the participatory web in which we find ourselves, and it will be revolutionary if an effective combination is achieved between the inclusion of semantic content in web pages and the use of artificial intelligence. Semantic codification will then become a fact when this is fully automated.

A collective digital intelligence is being constituted, and we are now faced with the prospect of multitudes of people with intelligences linked into the digital environment, interacting and having access to enormous amounts of information. The smart crowd emerges when communication technologies expand human cooperation talents. The technologies that are beginning to make the smart crowd possible are added to mobile
phones, social networks and all the mass of information that is transmitted through everyday devices. It is a global connection, linking individual intelligences. This astounding global connectivity could result in a budding future network of globally linked individual intelligences. As a hypothesis, we may contemplate the possibility that at present we are experiencing one of the greatest leaps in the evolution of our species. The connection of intelligences could well be the next step in the evolution of human intelligence, which seems like a process, because humanity has always had a great fascination with technological change. The new denominations of homo digitalis, homo connectatus or homo cyber already suggest a new conceptual stage in the era of humanity.

Accordingly, a global interconnection of information and semantic information is being generated, which may also produce an exponential pattern of technological growth, as Raymond Kurzweil explained. In the same way, technological advances in the digital world will give rise to machines that are increasingly powerful, numerous and cheap. Raymond Kurzweil elaborates further on this law in his book *The Singularity Is Near*, anticipating that the complexity of integrated circuits in computers will double every year, and suggesting that an exponential pattern of technological progress persists throughout human history.

Technological uniqueness is a future event predicting that technological progress and social change will accelerate. This event has been named by analogy with the gravitational singularity observed in black holes, where there is a point at which the rules of physics cease to be valid, and where the convergence towards infinite values makes it impossible to define a function. This pattern culminates in an unimaginable technological progress in the 21st century, which leads to the Singularity. Once having arrived at the point in which an intelligence superior to that of humans is created, we would enter a post-human stage. Kurzweil predicts that the first artificial intelligence will be built around a computer simulation of a human brain.

Forecasts like this point towards a total model of interaction with technology, which undoubtedly determines proposals for epistemologies. At present, a totalizing technological model of interaction has been reached that determines and floods the theoretical substrate of epistemologies, where at the same time our brain adapts to the digital. The structure of information, which has been mediated by orality and by writing, will now be by achieved by digital information and semantic interconnection. The medium has shaped the message, but now it seems to adjust not only the structure, but even to shape reality and, indeed, our own brain, as a digital technological paradigm emerges. In the 21st century, the agent of socio-economic change, scientific discipline par excellence, will be technology and digital information, which implies a certain technological determinism as a universal law.

Our postmodernity, with the digital network, has triggered revolutions in different
contexts, whether mediatic, scientific, epistemic, symbolic and organizational. Thus, the transition of modern epistemology to the postmodern and digital has been facilitated by the emergence of the technological paradigm of information. This new technological paradigm is a new way of thinking, doing and living in the digital environment. Consequently, the current epistemological configuration must address the treatment of a vastly-multiplied volume of multidisciplinary information from a heterogeneous digital environment.

Conclusions

The new digital materiality is transforming the material context, science itself, together with its organizational forms. Therefore, the new epistemological framework must be addressed, and knowledge organization must be shaped by an open, critical and transdisciplinary analysis.

At present, a totalizing technological model of interaction has been reached that determines and floods the substratum of epistemologies. The structure of information which has until recently been mediated by orality and writing, will now be conveyed by digital information, and semantic interconnection, with adaptation of our brain to the digital reality. The digital medium has shaped the message, but now it seems also to be shaping the structure, and what is more, it is already altering reality itself.

The WWW emerged as a technological aid for communication, as well as for commercial, business, economic, educational and military ends, among many others. Moreover, it became a space for discourse, and later imagination, as a space of identity, now constituting a fundamental part of our lives. For under every concept or construct a metaphor beats, and with its repeated use we forget that it is a metaphor. The metaphors that we habitually use shape our perception, to the moment when we no longer perceive them to be metaphors. It is not we who say them, it is they who tell us and conceptualize the world. In digital reality there beats a metaphor that has forgotten what it is, and that forgetfulness is what paradoxically gives it consistency. And it is precisely this occupation of the digital environment, where the digital metaphor thinks us, constitutes us, conforms and shapes us, that we cannot perceive. Consequently, there are innumerable and profound epistemological challenges to be addressed when it comes to organizing the knowledge of the digital age.

References

Capurro, Rafael. (2003) Epistemología y Ciencias de la Información. ENANCIB. Available at: http://www.capurro.de/enancib.htm#II.#II.


Karckhove, Derrick de (1997) Connected Intelligence: the arrival of the web society, Somerville House, USA.


Padrón, José. (2000). Los Procesos de Investigación (panorama histórico). Available at:


Sowa, J. F. Knowledge Representation, (2002). Available at:

Facet: itself a multifaceted concept

Abstract
In recent literature on knowledge organization, we have identified four ways of looking at facets. In this paper, we present these four ways of looking at and of exploiting facets, and we use elements related to our own research project to show that it is possible, and not necessarily harmful, for a single facet to assume a dual identity and to play more than one role in an information system.

1. Introduction
The concept of “facet” is more popular today than in S. R. Ranganathan’s own time. This does not mean, however, that a definitive characterization of this abstract concept has been established, or that all those who talk about or use facets are entirely clear as to the nature and function(s) of this handy device. In his article “Des classifications aux thésaurus: du bon usage des facettes”¹, Jacques Maniez (1999) adopted a Linguistics perspective to describe what he perceived as a significant weakness of the faceted approach. In his Colon Classification (CC), Ranganathan allocated concepts to one of five fundamental facets (Personality, Material, Energy, Space and Time) on the basis of their essential nature; but he also proposed suitable indexing formulas for each discipline, giving to each facet a specific function in the “sentence” thus created. Maniez implied that Ranganathan was not able to, or chose not to, dissociate the semantic (nature) and syntactic (function) dimensions of facets in knowledge organization (KO) and information representation. The ambiguity introduced by Ranganathan was maintained by the Classification Research Group (CRG) (1955) and in the many subsequent sets of facets inspired by the CRG proposal. The fusion/confusion of semantic and syntax is observed today in the diversity of definitions and applications in contemporary information systems.

Our research team has had to confront this ambiguity while looking for relevant facets around which to structure an indexing and retrieval system for legal decisions. An analysis of recent KO literature revealed four ways of looking at facets; these have influenced the selection and naming of facets in our project, as well as the system under development. In this paper, we present these four ways of looking at and of exploiting facets, and use elements of our own project to demonstrate that it is possible for a single facet to assume a dual identity and to play more than one role in a single information system.

¹ “From classification to thesauri: making good use of facets”.
2. Nature and role of the facet

The facet has been presented as one of the most significant theoretical developments of the 20th century in Information science. But what is exactly a facet?

The concept of facet is an abstraction and is more difficult to define than any concept representing an object that can be observed and physically manipulated. Maniez (1999, p. 262) remarked that, even after decades of usage, the meaning of the concept remained to be standardized, the term facet appearing both in highly specialized domains and in popular usage and being applied in the context of any grouping process in extremely diverse settings. Today, the technology that makes possible the sharing of responsibilities for information distribution has opened a path to an even wider interpretation of the nature and role of facets in information systems. In 2018, facets are used to refine a search in the library catalogue or to filter commercial goods online.

We have identified four conceptualizations of the facet in KO literature and through functional applications in contemporary information systems. Each conceptualization is presented below in the form of a dichotomy whose poles would appear at first glance to exclude one another; we will see that this is not always the case.

2.1. Process or product?

Facet, as a concept and as a term, is used in everyday language. In general language dictionaries, a facet is a dimension of a complex reality observed from a particular angle, as well as a characteristic common to all members of a class. KO specialists have used facets as criteria, as a basis for allocating objects or concepts to classes, making facets essential to the dividing process. Ranganathan had to use this interpretation when he determined that an entity belonged to the Personality rather than to the Matter facet.

But the facet is often also seen as the resulting class, the product of the dividing process. The facet has been defined as any grouping composed of entities sharing one or more characteristics. Ranganathan reused the term facet to name the set of isolates offered as potential manifestations of a fundamental category in his CC (Beghtol 1995, p. 197).

2.2. Nature or function?

In 2004, the Association française de normalisation (p. 103; our translation) defined facet as “a non-thematic category allowing for the categorization of a set of concepts on the basis of their essential nature or the perspective from which they are considered. For example: phenomenon, process, attribute, tool, etc.”. This definition suggests a second significant way of looking at the facet.

When reference is made to the nature, the essence of an object or a concept, the categorization process consists in allocating objects and concepts to a facet, giving them a meaning which will not vary with the context of use; this is what Ranganathan did in
the first stage of developing the CC, and most faceted classification structures have adopted this approach to using facet. Fundamental and mutually exclusive facets (for example, living being, physical object, attribute, activity, space, time) are capable of conveying the true nature of a concept.

The facet can also specify the role played by the concept, its function in a specific communication context, such as the analysis and representation of a complex subject. In this case, the allocating of a concept to a facet is determined by this concept’s relationship(s) with others rather than by its essential nature. A single concept, whatever its nature, can play various roles in subject representation, in the same way that a word can play different roles in different sentences. This conceptualization of the facet brings to mind linguistic techniques of discourse analysis, such as Charles J. Fillmore’s case grammar.

The CRG (1955) insisted on the importance of this functional dimension of the facet by demonstrating that the same entity did not belong to the same grouping (or facet) in different disciplinary environments; straw, for example, is an Entity in Botany, a Material in Construction, and a Product in Agriculture.

Maniez believes that Ranganathan’s indexing formulas assign a distinct function to each concept in the context of subject representation and notes that Ranganathan’s Personality, Matter and Energy facets are simultaneously fundamental and functional.

Both Mills (2004) and Vickery (2008) confirm that there exist two distinct types of facets: fundamental facets (living being, time, matter, etc.) and relational facets (agent, type, part, product, etc.).

2.3. Object or subject?

Maniez has suggested that the unresolved ambiguity surrounding the notion of facet has something to do with the type of entities that facets describe and/or categorize: physical objects such as commercial goods and library documents, or abstractions such as the components of a subject in a subject heading (1999, p. 261). Ranganathan establishes a clear link between facet and subject when he refers to the facet as a generic term designating any part of a complex subject (Maniez 1999, p. 253).

Should facets be used to describe an object (a document, for example) or to identify the components of an abstract subject? Although this distinction does not appear to reduce the efficiency of information systems, one must wonder if it is possible to use identical sets of facets to describe a physical object, represent a subject and structure the concepts of a discipline. Broughton (2001, p. 88) accepts all interpretations when she states that categories can be based on any attribute of documents/subjects (our emphasis) which it is useful to identify to facilitate indexing/classification and retrieval.

The association between facets and documents as objects is best illustrated by the frequent use of the term facet to name those elements of the bibliographic description
used as filters in a library catalogue. Authors’ names, original language, subject and other elements of the descriptive record are thus assimilated to facets. Gnoli (2008, p. 129) views the practice as a loose interpretation of the faceted approach, more common in North America than in Europe. Broughton (2006, p. 61) talks of non-semantic filters, but does not reject this conceptualization, which leads to a fourth interesting dichotomy.

2.4. Structure or navigation?

During the course of the 20th century, KO specialists have concentrated their efforts on the best use that could be made of facets to classify documents and organize collections. They continued to focus on hierarchies, notation and citation order, expending much energy in describing the design stages of faceted classification structures that would end up being quite similar to their predecessors.

While the function of classification is to group similar objects and subjects, that of indexing is to emphasize distinctions between similar subjects in order to facilitate a more targeted retrieval (Tudhope and Binding 2008, p. 217). Users must be able to isolate any aspect of a subject that is relevant to their information need; this presupposes that each component of the subject has been represented clearly and independently by the analyst (Vickery 2008, p. 148); Vickery suggests that each element thus represented is a facet.

While KO specialists were busy looking for the ideal classification system, other interested parties discovered Ranganathan and his faceted approach, seeing the potential of facets to describe and explore more or less structured digital contents. Facets came to be seen as essential keys for discovery. Broughton (2013, p. 742) remarks that very few existing faceted systems are based on a classification structure or use a controlled vocabulary; rather, facets are applied to structure the navigation interface and they play a critical role as search filters.

In the digital world, the need to organize and structure has lost much of its relevance. Nowadays, the first priority is to describe precisely all types of information objects. The second priority is the design of interfaces and search engines that will facilitate browsing, navigation, retrieval and presentation of results. A set of well-defined and functional facets specifies the dimensions on the basis of which an object or subject can be represented and retrieved.

Even if it is possible to discuss the concept of facet without reference to classification and classification systems, we believe that it is not essential to do so. The same facets (criteria) used to generate subsets of entities (classes) can also be used, conversely, to describe each entity without the necessity to preserve any link to its parent subset. For instance, sex, occupation, age, and residential address are facets used to group individuals; the same facets can be used to characterize any single individual. This explicit way of describing facilitates retrieval by criteria without any obligation to apply
syntactic rules (or citation order) or to know which class should be looked into (Hudon 2007, p. 152).

3. Selecting and naming relevant facets

The possible combinations of these conceptualizations of the facet (for example, facet as function used to describe a subject with a view to facilitating navigation, or facet used to organize a collection of objects according to physical characteristics) add complexity to the definition and theorizing of the concept.

In the framework of an ongoing project whose objective is to propose an innovative approach to the analysis and retrieval of legal decisions (Hudon and Cumyn 2017), our team has had to confront such complexity. The four ways of looking at facets have influenced the selection and naming of facets in our project, and determined to some extent the role(s) they will be playing in the new information system. We have observed that a single facet could assume a dual identity and play multiple roles within a single information system.

At the start of the project, we believed that it would be possible to identify a small set of fundamental facets which would allow us to organize concepts in the discipline of Law. But this needed to be done while taking into account the necessity to use functional facets allowing for the discrimination of concepts relating to facts and context, legal problem and applicable rules of Law. Diverging visions within our team led to a priority shift from structure to navigation.

We first established a bank of general concepts extracted from sources such as Moy’s (1968) classification, and Sweet and Maxwell’s (2010) *Legal Taxonomy*. To these concepts, we added specific notions identified through the process of indexing a representative sample of court decisions. A first set of fundamental facets was generated, which would allow for categorization on the basis of the essential nature of concepts. These fundamental facets were: Person, Process, Physical object, Attribute, Event and Space. We soon realized that it was too often impossible to allocate a concept to a single facet, and the six original facets were replaced by a restricted set of three fundamental facets: Person or relation (for example, husband), Thing, Event. Kaiser (as cited in Svenonius, 1978) and Bhattacharyya (as cited in Maniez, 1999) had indeed suggested that a smaller set of facets would solve the difficulty of creating mutually exclusive classes, but we discovered easily that it was far from being the case. Furthermore, a most important facet was now missing, that of Action (Ranganathan’s

---

2 *Conception et évaluation d’un nouveau modèle d’indexation de la documentation juridique* (Design and evaluation of a new indexing system for legal documents). Funded by the Canadian Humanities and Social Sciences Research Council for the period 2015-2020. Project leader: Professor Michelle Cumyn, Faculté de droit, Université Laval, Québec.

3 In his Systematic indexing system, Kaiser used three facets: Things (real or imaginary), Processes and Localities. Bhattacharyya built the POPSÍ system around four facets: Discipline, Entity, Action and Attribute.
Energy), which is almost always at the heart of a subject, or in our case, at the heart of the legal problem.

Moving away from fundamental facets, we tested a third set of facets, applicable to the Law domain. Each facet would now be considered functional, specifying the role of a concept in the statement of facts causing the legal problem, and in the description of consequences. A characterization of the legal discipline and practice suggested six functional facets: Subject (of Law), Action, Object, Context, Legal regime/rules, Consequences/Sanctions.

These six facets allowed for a precise description of non-legal elements in the representation of facts generating the legal problem and court case, as well as for the description of Law-related elements invoked by the judge in his or her decision. Following discussions and terminological adjustments, it was decided that the set of functional facets that would be used to index and retrieve legal decisions in our prototype database would be: Person, Action, Thing, Context, Legal regime and Remedy/Sanction were excluded from the set of functional facets, as they would not be extracted from the decision itself; they are now seen as classes to which decisions are allocated after analysis.

The facets and classes are being tested by indexers who have been provided with three autonomous controlled vocabularies. The first vocabulary is a structured list of descriptors, whose form and meaning have been standardized. Descriptors are associated to one of the four facets (Person, Action, Thing, Context) at the time of indexing only, each concept being given a function in the statement of facts leading to the court case. The second vocabulary is a classification system replicating the structure of laws and regulations applicable to the sub-domains represented in our prototype database. The third vocabulary is a controlled list of applicable remedies and sanctions.

In a system structured around functional facets, each concept is allowed to play different roles in different contexts, all the while preserving its true nature and essential characteristics. A concept can represent the action at the heart of a particular court case, but be an element of context in another case; for example, the accident is at the heart of the decision where the issue is one of responsibility, but it belongs to the context where the court has to judge the fairness of a sanction imposed to the worker who was late because of the accident. In our set of facets, Person and Action, and Person and Thing are mutually exclusive (a person can neither be an action nor a thing). Context is the most inclusive of all four facets since it will comprise concepts referring to persons, actions and things. Person and Action are simultaneously fundamental and functional facets, since they evoke both the nature and the function of a concept.
4. Conclusion

As a foundational element of classification systems, facets have traditionally been used to structure document collections, disciplines, even the whole of knowledge. In the digital information world, facets more often act as filters to facilitate navigation and retrieval, and to improve the presentation of search results. But the very nature of the concept of facet remains ambiguous.

Maniez has suggested that terminological refinements could clarify the contemporary discourse around the concept of facet (1999, p. 261). He proposed that the term facet always be associated with the qualifier “classificatory”, so that it is recognized as an Information science concept. He also distinguished two types of classificatory facets; categorial facets apply to any classification of concepts, regardless of domain, while structural facets describe the essential components of an entity or a subject. Maniez connects categorial facets with the paradigmatic axis and the structure of language, and structural facets with the syntagmatic axis and the discourse. There should be very few of the former but a great number of the latter, which will differ from one field of application to the others. Though intriguing, such terminological specifications are not sufficient to solve the original ambiguity as to the nature and function of facets. Furthermore, Maniez’s proposal reflects a conceptualization of the facet as a basis for classification, as a structuring device rather than as an aid to navigation and retrieval.

The term facet is widely used today by various groups of people concerned with the exchange of goods and of information; used in different contexts, it can designate a category, a class, a cluster, a characteristic, a criterion, an aspect, a filter; the term facet can designate the chronological dimension expressed by a subdivision in a subject heading as well as a non-essential attribute, its cost for example, of a product sold online. The ambiguity introduced by Ranganathan and maintained by the CRG and others authorizes such a wide appropriation of the concept; there seems to be no way around it.

It remains to be proven, however, whether this ambiguity has significant negative consequences on information representation and retrieval. Research is needed to examine the actual impact of the phenomenon on the quality of retrieval. Though facets and faceted analysis have been at the heart of numerous publications and scientific gatherings over the past 20 years, discussions seldom ventured beyond theoretical grounds. Applications of facets in operational retrieval systems, as abundant and popular as they may be, must be rigorously evaluated. As KO specialists, we need to demonstrate that the use of predefined facets in subject analysis and representation, as well as in the search and navigation interface, leads to quantitatively and qualitatively better search results, or minimally to results that are at least equivalent in relevance to those obtained in systems using descriptors or keywords, without syntax, to represent
the same subjects.

In our project, two stages of validation of the four functional facets and two classes guiding the indexer and the searcher (here the lawyer looking for one or more relevant court decisions) are scheduled. Usability studies will support our assessment of the usefulness of facets, and make possible a comparison of the relevance of decisions retrieved with or without the assistance of those same facets.

References


Feasibility of implementing PRESSoo model in organizing Persian serials

Abstract

Purpose: The purpose of this study was to identify the degree of compatibility of Persian serials with the PRESSoo model.

Methods: This paper describes a survey study conducted on a population consisting of 90 Library and Information Science (LIS) serials from the National Library of Iran. Firstly, a mapping table was produced consisting of ISSN data elements and PRESSoo classes, as published in the last PRESSoo report. Secondly, this mapping table was adapted to fit UNIMARC fields, and used to determine to what extent Persian LIS serials are compatible with the PRESSoo model. In order to be sure about the accuracy of the checklist, reference was made to ISSN manual 2015 in some cases.

Results: Results of this study showed that bibliographic records for Persian LIS serials are compatible with 5 classes and 18 properties of PRESSoo. The reason for this low compatibility is due to library software problems, leading to inattention to some key UNIMARC fields. Some important classes are absent in the studied records. In order to have adequate catalogs, some important classes such as end of publication, storage units for cumulative works, supplementary works, and their properties should be added to Persian serial catalogs.

Value: PRESSoo is a new conceptual model for describing important elements of serials and continuing resources. In 2017, this model was designated as an IFLA standard. However, this model has not been used in Iran. This study has analysed the feasibility of implementing the PRESSoo model for organizing Persian serials.

Background

In order to overcome the problems of library catalogs, IFLA and other international organizations developed new models. One of the most important of these models is FRBR, which was created in 1998.

FRBR obeys entity-relationship formalism. Oliver (2017) notes that the "decision to use entity-relationship modelling was also an important step in changing the library’s perspective of bibliographic information, from strings of information embedded in paragraphs to data that can be handled efficiently in digital environments. Entity-relationship modelling has to be understood outside the traditional library record structure. Entities not held together by being in the same “record”. Libraries working in a MARC environment are accustomed to pieces of information being held together within the shell of a bibliographic or authority record. There is no record in the entity-relationship model. The entities are held together in a meaningful way solely through the relationships".
Problem statement

FRBR is used for modelling different resources, but serials are always absent from these studies, regarded as being too problematic. One reason why serials are not usually seen as convenient for implementing FRBR is that they are aggregate works. Each serial is composed of smaller independent works that are intellectual works in their own right. This leads to a modeling challenge when applying the FRBR conceptual model. (Ballegooie and Borie 2014). The FRBR model appears not to be flexible enough to accommodate serial relationships and, above all, its static nature can lead to inconsistent assertions when applied to serials (Le Boeuf 2014).

Definition of works and their manifestation in serials, changes in title, publisher, frequency of serial, among other issues, require more attributes and relationships. As a result, FRBR is not well suited to the dynamic nature of serials.

In Iran, catalogs are produced according to Anglo-American cataloging rules. Because of such rules and the flat environment of MARC, the position of each work and its relation to similar works in the bibliographic universe is not explicitly stated, and existing catalogs do not provide users with the links between works.

There are many additional problems with serials. Few libraries completely catalog their periodicals. The lack of consistency as regards receiving periodicals and their absence from libraries is one of the main reasons for neglecting their importance and not organizing them in a coherent and complete way.

However, the benefits of cataloging such resources should not be ignored. One such benefit is that it enables more access points to a publication, while also facilitating information searches in publications.

PRESSoo seeks to introduce a new concept of bibliographic information applicable to continuing resources. This model is based on the high-level concepts of FRBR and attempts to provide further details and attributes for such resources. Accordingly, PRESSoo is used as a valid ontology to express the semantics and relations of continuing resources. However, these benefits require conducting research in order to implement this model in serials and, identify appropriate models for these resources.

Functional requirements for object-oriented bibliographic records

The problems of implementing FRBR in serials were stated during a harmonization meeting held between representatives of ISBD, RDA and ISSN in 2011 (Le Boeuf 2014).

These problems were discussed by different FRBR experts (Françoise Leresche and Patrick Le Boeuf from the Bibliothèque nationale de France (BnF), Philippe Le Pape from ABES, and Gordon Dunsire (IFLA consultant). During a technical meeting in Paris, in April 2012, held in conjunction with the General Assembly of the ISSN Network. Patrick Le Boeuf’s presentation of FRBRoo (FRBR model based on object-
oriented modelling) showed that this model could be a possible solution (Bequet, Howlett and Willer 2015).

In order to understand FRBRoo, one should be familiar with the CIDOC Conceptual Reference Model. CIDOC CRM is an ontology to facilitate the interchange of cultural heritage, developed by the International Council of Museums (ICOM). Work on CIDOC CRM started in 1996 with the primary role of enabling information exchange and integration between heterogeneous sources of cultural heritage information. It aims at providing the semantic definitions and clarifications needed to transform disparate entities (CIDOC CRM final report 2017).

Because the contexts in which FRBR and CIDOC CRM were designed could not have been more different, libraries and museums fall under the common umbrella of “cultural heritage institutions”. By the end of the 1990s, experts started to talk about “ALM convergence”. Since archives, libraries and museums are cultural heritage institutions, they work together. Thus, in 2003, a “harmonization” between FRBR and CIDOC CRM was formed. The outcome of this endeavor was published in 2009 as version 1.0 of FRBRoo (Le Boeuf 2012).

FRBRoo is a reformulation of FRBR and CIDOC CRM. FRBRoo appeared to provide useful classes and properties for modeling serials. This model is an event-based and object-oriented model, it means that it is able to better express serials, and it offers classes and properties for modeling seriality. However, FRBRoo deals with serials at very general levels and it was not suitable for serials (PRESSoo final report 2017).

At the end of 2012, the ISSN International Center and Bibliothèque Nationale de France agreed to establish a working group in order to develop an extension of FRBRoo that would be devoted to serials and continuing resources. In January 2013, the working group started to draft PRESSoo, whose Version 1.0, endorsed by the FRBR Review Group, was released in June 2014.

PRESSoo

PRESSoo is an extension of the FRBRoo and CIDOC CRM models. This model is a formal ontology intended to capture and represent the underlying semantics of bibliographic information relating to continuing resources, and more specifically to serials. Cataloguing of continuing resources is always a major difficulty in libraries due to the modelling challenges of these resources. Continuing resources may undergo many changes through time. Their bibliographic information, such as publisher, country of publication and publication frequency, may change. These problems, make it harder to catalog a serial than a monograph. In the final PRESSoo report (2017) it was stated that: "The main difference between cataloguing a monograph and cataloguing a serial could be expressed as follows: when you catalogue a monograph, you make statements about the past; when you catalogue a serial, you both make
statements about the past and assumptions about the future”.

**Purposes of PRESSoo**

- To express the common, implicit conceptualization that underlies bibliographic description of continuing resources;
- To formalize bibliographic information as economically and appropriately as possible as a set of triples;
- To provide a complete set of classes and properties that may be used in conjunction with other bibliographic ontologies to describe continuing resources;
- To provide all solutions to precisely express any kind of bibliographic information that is considered worthy of publication.

Implementers of PRESSoo are able to choose object-centric or event-centric models, demonstrating the model’s flexibility. PRESSOO may be used, in conjunction with CIDOC-CRM and FRBROO, as the main ontology to publish a linked dataset. (PRESSoo final report 2017).

Le Boeuf (2015) outlined differences between object-oriented models and former conceptual models. These differences are:

- Object-oriented models do not mention any user task. This is the first difference between object-oriented approaches with FRBR, FRAD AND FARSAD;
- In object-oriented models, the term 'class' is used to refer to the basic elements in the bibliographic universe. The things that belong to a given class are called instances of that class;
- Classes are declared to be subclasses of FRBRoo and CIDOC CRM. Any instance of a class that is declared to be a subclass of another class is automatically also an instance of that other class. This notion is called inheritance;
- There are simply no attributes at all. Many notions that were merely declared as entity attributes in FRBR/FRAD/FRSAD are now modelled as classes in their own right, since it emerged that they were more complex than initially assumed. The most striking example is that of Date – an attribute that is declared for several entities in FRBR/FRAD/FRSAD. The initial assumption was that a date is a mere literal term associated with, say, a work or a manifestation;
- In object-oriented models, relationships are called properties.

**Classes and properties of PRESSoo**

PRESSoo classes are identified by a letter Z, a number and a noun phrase. PRESSoo has 46 properties, which are held between a domain and a range. A domain declares the class for which the property is defined, and a range defines the class which the property indicates (Riva, Doerr and Zumer 2008). Classes borrowed from CIDOC CRM are identified by a letter E, a number and a noun phrase. In addition, classes borrowed from
FRBRoo are identified by a letter F, a number and a noun phrase. Table 1 shows classes of PRESSoo and their super classes.

Table 1: PRESSoo classes

<table>
<thead>
<tr>
<th>PRESSoo classes</th>
<th>Subclasses of</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z1 Serial Transformation</td>
<td>F27 Work Conception</td>
<td>Instances of continuation, replacement, split and merger.</td>
</tr>
<tr>
<td>Z2 Absorption</td>
<td>F27 Work Conception</td>
<td>Activities absorbed by preexisting instances of F18 Serial Work.</td>
</tr>
<tr>
<td>Z3 Separation</td>
<td>F27 Work Conception</td>
<td>Activities that come into being as the ‘offspring’ of pre-existing instances of F18 Serial Work that continue to be published after the separation has occurred.</td>
</tr>
<tr>
<td>Z4 Temporary Substitution</td>
<td>F30 Publication Event</td>
<td>Publication of instances of F18 Serial Work as a temporary replacement for another instance of F18 Serial Work</td>
</tr>
<tr>
<td>Z5 Issuing Rule Change</td>
<td>E7 Activity</td>
<td>Change arising from Issuing Rule</td>
</tr>
<tr>
<td>Z6 Starting of Publication</td>
<td>F30 Publication Event</td>
<td>Publication of the first issue of a given continuing resource.</td>
</tr>
<tr>
<td>Z7 Ending of Publication</td>
<td>F30 Publication Event</td>
<td>Publication of the latest known issue of a given continuing resource.</td>
</tr>
<tr>
<td>Z8 Metadata Management</td>
<td>E7 Activity</td>
<td>Activities of ISSN centers, involving responsibility for the metadata relating to a continuing resource</td>
</tr>
<tr>
<td>Z9 Storage Unit</td>
<td>F4 Manifestation Singleton</td>
<td>unique combinations of instances of E18 Physical Thing that are bound together</td>
</tr>
<tr>
<td>Z10 Sequencing Pattern</td>
<td>E55 Type</td>
<td>Anticipated formats to be used in designating volumes/issues, etc. and/or dates for the individual units of the serial (e.g. Volume ..., number ...)</td>
</tr>
<tr>
<td>Z11 URL</td>
<td>E51 Contact Point</td>
<td>Identification of servers from which digital files can be obtained online</td>
</tr>
<tr>
<td>Z12 Issuing Rule</td>
<td>E29 Design or Procedure</td>
<td>Specification of the issuing policy to be followed at a given point in time for instances of F18 Serial Work.</td>
</tr>
<tr>
<td>Z13 Monograph</td>
<td>F19 Publication Work</td>
<td>Instances planned to result in an instance of F24 Publication Expression which should either be completed as a single part or completed within a finite (and predetermined) number of parts, as opposed to instances of F18 Serial Work.</td>
</tr>
<tr>
<td>Z14 Storage Unit Creation</td>
<td>F28 Expression Creation</td>
<td>Activities through which instances of Z9 Storage Unit are produced.</td>
</tr>
</tbody>
</table>
Continuing resources also display complex relationships: they may originate from continuations, splits or mergers with other continuing resources; they may be (local, linguistic, special) editions of other publications – and these relationships themselves may evolve in the lifetime of the resources.

PRESSoo takes the notion of E29 Design or Procedure, which in FRBRoo provides complete coverage of issuing policies for continuing resources in all their aspects, specializing it into the Z12 Issuing Rule class, which covers only one of all the aspects of the overall policy of a continuing resource. In FRBRoo all instances of serial changes are modelled as part of just one instance of E29 Design or Procedure; in PRESSoo, the same changes are modelled through different instances of the Z12 Issuing Rule. This makes it easier to account for the fact that not all aspects of a serial may change at the same time; modelling the overall issuing policy as a block requires duplicating all the characteristics that remain unchanged every time a single characteristic (e.g. frequency) is modified. Class Z5 Issuing Rule Change is used to specify modifications made by the publisher (Le Boeuf 2014).

Methodology
This paper describes a survey study. To gather the required data, firstly, a mapping table was produced consisting of ISSN data elements and PRESSoo classes, as published in the last PRESSoo report. Secondly, this mapping table was adapted to UNIMARC fields identified by the ISSN Manual before was then being used to determine the degree of compatibility between Persian LIS serials and the PRESSoo model.

Population
The population of this study consists of 90 Library and Information Science serials from the National Library of Iran which are cataloged in RASA. This software is one of the most important library software packages in IRAN, produced by Pars Azarakhsh. RASA is designed especially for the National Library of Iran and, based on IRANMARC, which is in turn based on UNIMARC.

The pie chart below shows the different types of serials studied in this research.
Among the records studied, there were 41 newsletters, 40 scientific journals, 4 internal newsletters of organizations, 3 newspapers and 2 special issues.

To identify the appropriate model for Persian serials, events occurring in the records were determined by checklist, including: title changes, changes of publisher, change of publication frequency. The table below shows changes that occurred in the records studied.

Chart 2 shows that the main change occurring in serials is title changes. The main title changes are due to continuation, replacement and splits. The frequency of title changes is shown in table 2.

Results

The necessary information was gathered from studying the checklist based on PRESSoo classes and IRANMARC fields. The results are shown in the table below.
Table 2: PRESSoo classes and their compatibility with the records studied

<table>
<thead>
<tr>
<th>Classes</th>
<th>Frequency of records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z1 serial transformation</td>
<td></td>
</tr>
<tr>
<td>Continuation</td>
<td>12</td>
</tr>
<tr>
<td>Replacement</td>
<td>17</td>
</tr>
<tr>
<td>Split</td>
<td>1</td>
</tr>
<tr>
<td>Merger</td>
<td>-</td>
</tr>
<tr>
<td>Z2 Absorption</td>
<td>-</td>
</tr>
<tr>
<td>Z3 Separation</td>
<td>-</td>
</tr>
<tr>
<td>Z4 Temporary Substitution</td>
<td>-</td>
</tr>
<tr>
<td>Z5 Issuing Rule Change</td>
<td>-</td>
</tr>
<tr>
<td>Z6 Starting of Publication</td>
<td>-</td>
</tr>
<tr>
<td>Z7 Ending of Publication</td>
<td>-</td>
</tr>
<tr>
<td>Z8 Metadata Management</td>
<td>-</td>
</tr>
<tr>
<td>Z9 Storage Unit</td>
<td>-</td>
</tr>
<tr>
<td>Z10 Sequencing Pattern</td>
<td>-</td>
</tr>
<tr>
<td>Z11 URL</td>
<td>-</td>
</tr>
<tr>
<td>Z12 Issuing Rule</td>
<td>-</td>
</tr>
<tr>
<td>Z13 Monograph</td>
<td>-</td>
</tr>
<tr>
<td>Z14 Storage Unit Creation</td>
<td>-</td>
</tr>
</tbody>
</table>

Chart 3: Compatibility between PRESSoo classes and serial records

As shown in table 2, 12 continuations, 17 replacements and 1 split were recognized in the studied. The data gathered shows that an appropriate model must pay attention to the kind of changes that can occur in serials and establish a corresponding relationship between two such serial works. Any change that occurred in the serial is modelled through the Z5 Issuing Rule Change class. Another important notion of PRESSoo is its properties. The compatibility between serial records PRESSoo properties are very important since the relationship between two instances of classes is shown by a property.
Table 3: PESSoo properties already used in records studied

<table>
<thead>
<tr>
<th>Property</th>
<th>Domain of property</th>
<th>Range of property</th>
<th>Number of compatible records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1 provided a continuation to (was continued through)</td>
<td>Z1 serial transformation</td>
<td>F18 serial work</td>
<td>6</td>
</tr>
<tr>
<td>Y2 initiated as continuation (was initiated as continuation through)</td>
<td>Z1 serial transformation</td>
<td>F18 serial work</td>
<td>4</td>
</tr>
<tr>
<td>Y3 provided a replacement to (was replaced through)</td>
<td>Z1 serial transformation</td>
<td>F18 serial work</td>
<td>10</td>
</tr>
<tr>
<td>Y4 initiated as replacement (was initiated as a replacement through)</td>
<td>Z1 serial transformation</td>
<td>F18 serial work</td>
<td>7</td>
</tr>
<tr>
<td>Y6 initiated (resulting from a split)</td>
<td>Z1 serial transformation</td>
<td>F18 serial work</td>
<td>1</td>
</tr>
<tr>
<td>Y15 replaced (was replaced through)</td>
<td>Issuing Rule Change</td>
<td>Z12 Issuing Rule</td>
<td>17</td>
</tr>
<tr>
<td>Y16 replaced with (was introduced through)</td>
<td>Issuing Rule Change</td>
<td>Z12 Issuing Rule</td>
<td>12</td>
</tr>
<tr>
<td>Y17 launched (was launched through)</td>
<td>Starting of publication</td>
<td>F18 Serial Work</td>
<td>90</td>
</tr>
<tr>
<td>Y21 foresees use of language (is language foreseen in)</td>
<td>Z12 Issuing Rule</td>
<td>E56 Language</td>
<td>90</td>
</tr>
<tr>
<td>Y23 foresees dimension (is dimension foreseen in)</td>
<td>Z12 Issuing Rule</td>
<td>E54 Dimension</td>
<td>10</td>
</tr>
<tr>
<td>Y24 foresees use of title (is title foreseen in)</td>
<td>Z12 Issuing Rule</td>
<td>E35 title</td>
<td>90</td>
</tr>
<tr>
<td>Y28 foresees URL (is URL foreseen in)</td>
<td>Z12 Issuing Rule</td>
<td>Z11 URL</td>
<td>16</td>
</tr>
<tr>
<td>Y29 evolved into (continues)</td>
<td>F18 serial work</td>
<td>F18 Serial Work</td>
<td></td>
</tr>
<tr>
<td>Y31 was superseded by (superseded)</td>
<td>F18 serial work</td>
<td>F18 Serial Work</td>
<td>10</td>
</tr>
<tr>
<td>Y32 was split into (resulted from splitting)</td>
<td>F18 serial work</td>
<td>F18 Serial Work</td>
<td>1</td>
</tr>
<tr>
<td>Y37 has former or current issuing rule (is former or current issuing rule of)</td>
<td>F18 Serial Work</td>
<td>Z12 Issuing Rule</td>
<td>32</td>
</tr>
<tr>
<td>Y38 has current issuing rule (is current issuing rule of)</td>
<td>F18 Serial Work</td>
<td>Z12 Issuing Rule</td>
<td>90</td>
</tr>
<tr>
<td>Y44 foresees topic (is topic foreseen in)</td>
<td>Z12 Issuing Rule</td>
<td>E1 CRM entitie</td>
<td>90</td>
</tr>
</tbody>
</table>
All the properties used in PRESSoo have a domain and a range. As shown in table 3, 15 properties of the PRESSoo model are used in the records studied. The highest compatibility between properties and serial records are found in the title, publication date and topic fields of the IRANMARC records used in RASA. In the records studied just 30 changes in Issuing Rules were identified, namely: 28 title changes, 1 publisher's name change and 1 frequency change. The Z11 URL class does not have any property but, the Z12 Issuing Rule has a property for associating a URL to a continuing resource, specifically Y28 foresees URL (is URL foreseen in). As table 3 shows, the domain or range of some properties are referred to FRBRoo or CIDOC CRM classes since PRESSoo is an extension of FRBRoo and CIDOC CRM models. Therefore, attention should be paid to super classes of PRESSoo classes so that they can be used in a new model.

Conclusion
PRESSoo is a new model for describing the semantics underlying serials and continuing resources. Work on PRESSoo started in 2013, and the first version of this model was published in 2014. In 2017, PRESSoo was designated as an IFLA standard.

In order to test the PRESSoo model, ISSN Centre with the support of UNESCO, developed ROAD, the Directory of Open Access scholarly Resources. This service, launched in December 2013, provides free access to a subset of the ISSN Register, currently gathering around 15,000 bibliographic records.

As Oury (2017) states, PRESSoo may appear to be a complex ontology. On one hand, it inherits its complexity from the CIDOC-CRM, which deals with a huge domain. This complexity also arises from the ambitious scope of FRBRoo; offering a model able to describe any situation occurring in the bibliographic domain, unlike FRBRER which cannot always cover all cases. However, such complexity should be viewed as a strength; the richness of this model offers many usage opportunities.

The records of Persian serials in the field of LIS are compatible with 5 classes and 18 properties of the PRESSoo model. The reasons for this low level of compatibility
are: library software problems and inattention to some important UNIMARC fields. Some important classes are absent in the records studied. For example, there was no end of publication field in the records. Z7 ending of publication is an important class for identifying ceased serials. In order to have adequate catalogs. Some important classes like end of publication, storage units for cumulative works, supplement works, and their properties should be added to the Persian serial catalogs.

Iranian librarians need to become more familiar with these models to identify an adequate model for organizing serials and develop current standards. In the MARC environment, catalogers usually have no alternative for describing changes in serial resources. A linked data environment may offer greater flexibility in representing changing information. By using conceptual models, users can much more quickly receive the best responses, and the combination of library catalogs with ontologies can become a possibility.

Suggestions

As stated in this research, PRESSoo is an extension of the FRBRoo model. In order to identify the appropriate model, it is important to pay attention to FRBRoo and CIDIC CRM classes. It is suggested that further studies be conducted to determine the relationship between these three models.

This study has identified the feasibility of using the PRESSoo model in Persian LIS serials with RASA software. It suggests that further studies are undertaken to examine PRESSoo in other Iranian Library software packages to identify the best context and software for using the PRESSoo model.

References


Foundations and methods for Knowledge Organization in European iSchools courses

Abstract
The teaching and learning of information and knowledge organization is essential when training an information professional. In order to formulate a common teaching strategy directed to information studies in the European area, this paper presents the preliminary results of an ongoing exploratory study related to how the foundations and methods of knowledge organization are taught in European iSchools. Despite the limited information in English regarding undergraduate studies, it is possible to explore a common framework of training in knowledge organization in both bachelor’s and master’s degrees.

Introduction
The teaching and learning of information and knowledge organization is essential when training an information professional. However, in order to reach consensus, a clarification of these concepts is necessary. Among academic courses developed in the field of information science, the options regarding the designations of curricular units/modules related to these subjects are not the same in most schools, and it is therefore necessary to explain our understanding of their meaning. The use of the two terms, information and knowledge, is due to the fact that they are often applied randomly, with no obvious distinction between their underlying meanings. On the other hand, curricular plans do show a differentiation, but that does not mean that these two branches of study should not both be considered as important constituents of the learning process in information science. The precise distinction between the terms needs to be further elaborated, as acknowledged by Pando and Almeida (2015), who say that terminological analysis still reveals no clear pattern.

It is thus clear that, underlying this study, there are both theoretical and practical teaching methods regarding ways of processing information with a view to its retrieval. This implies, therefore, a process of descriptive cataloging and subject indexing; that is, instruction in how to attribute descriptive metadata and authority control access points.

Research related to the teaching of information and knowledge organization are scarce and, in general, are related to the most targeted courses/modules for the technical treatment of bibliographic information. Nevertheless, such studies play a valuable role in determining the importance of teaching the aforementioned subjects in the scope of information science (Joudrey and McGinnis 2014; Hudon 2010; Pattuelli 2010; Joudrey 2008; Davis 2008; Bowman 2006; Ruiz-Perez and López-Cózar 2006; Taylor 2004; Taylor and Joudrey 2002).
The decision to study iSchool courses was taken due to the increasing growth of this network of university schools in the field of information, which is currently expanding into Asia, Europe and North America. Despite the differences existing between iSchools structures, and the courses and type of training provided, they share a common standpoint, which is an interest in the relations between information, people and technologies.

Objectives and methodology

In order to formulate a common teaching strategy, directed to information science studies in the European area, this paper presents the preliminary results of an ongoing exploratory study related to how the foundations and methods for knowledge organization are taught in European iSchools.

The study included undergraduate (bachelor) and graduate (master) courses for the following indicators: course designation, module designations, contents, fundamental methods and approaches, recommended bibliography, teaching methods, student assessment components and the statute of the module: compulsory or elective.

The data were collected and / or confirmed in September 2017 and again in January 2018 in order to obtain an updated picture of training provision. It was found that several courses had undergone some changes in the structure of their curriculum, as well as in the provision of curricular units/modules. This was due, in part, to the restructuring of schools where teaching is given, and also to mergers with other courses, with a view to broadening the scientific basis of the courses.

The websites of the 26 schools mentioned in the European directory of iSchools were consulted (available at http://ischools.org/stories/regions/european-ischools/european-directory/). It was verified that most of the undergraduate courses are taught in the country's official language and the contents of curricula and modules are not available in English. On the other hand, undergraduate study cycles are undergoing reorganization of their study plans in order to provide wider education opportunities and to allow several options in connection with master’s degrees. Nevertheless, data available in English or Spanish were collected. Information about master’s degrees is mostly available in English and international student exchange is expected in several courses.

Results and discussion

Data were collected from the websites of 9 iSchools, specifically regarding 3 undergraduate courses and 9 master’s courses. Comparing the available information, it can be seen that a similar approach is used in subjects dealing with knowledge organization, although corresponding modules are differently distributed and designated, with some courses, for example, offering several modules dedicated to knowledge organization. The curricular structure of the courses which offer these
subjects is also similar; in other words, providing education in the field of information science as distinct from computer science or management, even though courses have very diverse designations. They may range from Information Management, through Information Science, to Management of Information and Digital Contents.

Modules dedicated to knowledge organization in the undergraduate courses are named as follows: Description of Information Sources, Descriptive Metadata, Organization of Information, Indexing and Classification, Knowledge Organization and Representation, and also, Authority Control and Subject Indexing.

When comparing credits given to these modules, it is possible to conclude that the modules are considered important subjects, since the number of European Credit Transfer System (ECTS) goes from 5 to 9 credits. Almost all modules are compulsory and the electives (Description of Information Sources, and Indexing and Classification) are preceded by a compulsory module, which prepares students with the essential contents regarding cataloging and subject indexing. These modules have theoretical and practical classes, and their contents can be systematized according to the following items: basic concepts about organization of information, metadata schemes, international standards, controlled vocabularies, classification systems, automation of authority control, and collaborative representation of knowledge. Some basic bibliography is common, but each school mentions several titles in the country's official language. Assessment methods in all modules include assignments and final exams.

The educational provision of master’s courses in European iSchools is very extensive and the designations reflect the schools’ options in relation to the scientific specificity to be attended. Regarding the approaches to knowledge organization, the courses which include this subject are variously named as Digital Curation, Digital Library Management, Information Management, Information Science, Information Science and Cultural Communication, Librarianship, Libraries, Archives and Digital Continuity, Library and Information Services Management, and Library and Information Studies. Some schools offer the same modules related to knowledge organization in different courses, highlighting the importance of this subject. In most courses, information is less detailed than in undergraduate courses, explaining in general the topics included in the contents. ECTS assigned to the modules vary widely from one course to another and may range from between 6 and 15 credits, and the statute of the modules is almost always compulsory. As far as the designation of modules is concerned, there is no substantial difference between them, particularly when compared with undergraduate courses: Archives and Records Management in the Digital Environment, Cataloguing and Classification, Content Analysis and Indexing, Information Organization, Information Organization and Access, Knowledge Organization, Organization and Analysis of Information, and, lastly, Organization of Information, Cataloguing and Metadata. Module content and learning outcomes are
very similar to the items mentioned in the undergraduate courses; that is, preparation of students to organize information and knowledge in various domains, including business, web, archives, libraries, and museums. This fact can probably be explained by the circumstance that most schools offer only master’s courses. Nevertheless, teaching methods may contribute to a distinct approach to the subject. In fact, graduate courses include seminars, laboratory work, discussions, and workshops. The most common method of assessment is by assignments, followed by final exams. The selected bibliography is mentioned in a limited number of modules, but it is almost in English and some titles are common to several modules. In Appendix I there is a list of all references included in the bibliography section of module information (in English). This can contribute to the definition of a common list of compulsory bibliography to be applied in all courses, although it was identified that some references need to be revised to include updated editions.

Conclusions

Despite the limited amount of information in English regarding undergraduate studies, it is possible to explore a common framework of training in knowledge organization, both in bachelor’s and in master’s degrees. This common framework may enable the establishment of synergies between the targeted schools with a view to presenting a common or complementary training provision through online seminars or workshops, which could stimulate contact between students. The sharing of resources and teaching components, such as the provision of educational materials, can be enhanced, allowing closer contact between pedagogical practices and the eventual creation of collaborative research projects.

The limitation of this study is that data collection and interpretation is based on information available in English or Spanish from the course websites. Some important information regarding the courses offered at other schools would be relevant for the establishment of a common strategy for teaching knowledge organization. These issues will be explored in further stages of this research.

References


Appendix I

References mentioned in the modules

Guangyuan Sun, Christopher S. G. Khoo

A Framework to represent variables and values in Social Science research data sets to support data curation and reuse

Abstract
Research data management (RDM) initiatives have been increasingly adopted by many disciplines. Data curation is part of RDM and aims to support data reuse. This paper describes our framework for the representation and organization of social science research quantitative data sets generated from survey questionnaires to support data reuse. The framework comprises three levels: Ontology Level, Conceptual Description Level, Physical Description Level. The Conceptual Description Level and the Physical Description Level are directly derived from and represent the structure and content of physical data files in any format (e.g., spss). The Ontology Level is derived from questionnaires related to data set files, and is designed to support the assignment of semantics to data set and to deal with issues of semantic ambiguity. Finally, this paper discusses the characteristics of our framework from two aspects: 1) in comparison with the Data Documentation Initiative standard; 2) our high-level concept \textit{valueType} and its benefit for the representation of social science data sets to support data reuse.

Introduction
An increasing number of universities are embarking on research data management (RDM) initiatives, including encouraging faculty members to develop data management plans, and building institutional data repositories to support the preservation and reuse of research data from their faculty members and research staff (Antell, Foote, Turner and Shults 2014; Neylon 2017).

Data curation is included in RDM, and considered as one part of the research data services provided by university libraries. It is the systematic archiving and management of research data with the intention of making it available for use by other researchers beyond the lifespan and purpose of the project for which the data was collected (Rusbridge \textit{et al.} 2005; National Science Foundation 2005; Digital Curation Centre, 2018).

The main aim of data curation is to support future data reuse. However, the definition of data reuse in existing literature remains unclear. We take the position that as long as data is used for purposes other than that for which it was originally collected, data is reused. We think when reusing data, users need to make the conceptual effort of mentally bridging the conceptual gap between the meaning of terms used in the original data files, and the meaning of these terms in their knowledge base (Sun and Khoo 2016).

In the past, data curation initiatives have focused on scientific communities, especially in the bioscience area (Howe \textit{et al.} 2008). Research studies on social science data curation are relatively lagging behind. An increasing amount of social science data relating to people, organizations and economic activities is becoming publicly
available. Big social science data repositories that host major social science data sets, such as the Inter-university Consortium for Political and Social Research and the UK Data Archive, have already been set up. Governments are setting up national data repositories to make socio-economic-demographic data collected by government agencies available.

This study aims to investigate the curation of social science research data to support data reuse. Social science research data can generally be divided into two categories: quantitative (e.g., tabular data sets from survey questionnaires or experiments) and qualitative (e.g., textual data, audio, video, and images). Different types of data create different reuse challenges for users. Based on Whitmire, Boock and Sutton (2015)’s study about variability of research data management, quantitative data (e.g., spreadsheets, delimited text, SPSS) are the most commonly created type of data across twelve scientific and social science disciplines. In the hope that our study can shed light on the representation of quantitative data sets beyond the social science community, this paper targets quantitative data sets generated from social science survey questionnaires.

The challenges of reusing quantitative social science data sets come from the substantial intellectual effort users need to make to understand the meaning of the variables and values in the data sets:
- Different researchers may differently name variables that refer to the same or similar concepts. Alternatively, variables with the same name may refer to different concepts;
- Relationships between variables may not be readily apparent. Users may need to read accompanying documentation closely to understand them;
- The unit of measure or scale for numeric values may not be stored with the data set;
- The semantics of categorical values may not be obvious to the user.

Thus, the structure and semantics of variables and variable values in data sets need to be represented accurately. This can be achieved using knowledge organization and representation techniques. We have chosen to use the technique of ontology to deal with the above challenges. This paper describes the framework that we have developed using ontology for the representation of data sets and discusses how its characteristics support social scientists in reusing data sets in comparison with the well-known and widely-adopted Data Documentation Initiative (DDI).

**Methods**

We have collected a sample of 507 data records (i.e., data sets and associated survey questionnaires) from two well-known social science data repositories (Inter-university Consortium for Political and Social Research and UK Data Archive), and analysed their
content and structure to harvest requirements for accurately and comprehensively representing data sets.

ICPSR contained 9,970 data records as of May 2017, categorized into 19 topics. We systematically sampled 13 data records from each topic: 10 samples are used for building the ontology, and another three samples are used for testing the completeness and comprehensiveness of the ontology. The same sampling procedure was applied to the UK Data Archive, which contained 7,890 data records, categorized into 20 topics at the time of sampling (June 2017).

This study implements the framework in computers as text-based files using the ontology representation language OWL (Web Ontology Language) Level 2. The choice is obvious as OWL is used in the Internet environment to support the semantic web and linked data applications. To support the development of the ontology, we used the ontology editor software TopBraid Composer Standard Edition.

Our ontology framework

This section describes the framework that we developed (Figure 1) for modelling data sets. Briefly, the framework has three levels (Ontology level, Conceptual Description Level, and Physical Description Level), each level being sub-divided. The Conceptual Description Level and the Physical Description Level are directly derived from and represent the structure and content of physical data files in any format (e.g., spss). The Ontology Level is derived from questionnaires related to data set files and is designed to support the assignment of semantics to data set and to deal with issues of semantic ambiguity. The rest of this section introduces each level in detail.

Physical description level

Adopting the terminology of relational database theory (Date 2000), a data set is made up of one or more relation values (i.e. data tables), each with two parts – a heading and a body:

- The heading is a set of \{attribute-name (i.e. column-name): type-name (i.e. datatype)\} pairs, such as \{“Gender”:“Integer”\}. The type-name component is to represent the fact that each attribute is associated with a simple datatype (e.g., integer, character string, and date), meaning that the values in the attribute are members of a datatype.

- The body is a set of tuples (i.e. rows of records) that conform to the heading. Each tuple contains exactly one value for each attribute. Tuples are distinguished by having unique values in the primary key header (i.e. unique identifier), such as \{“ID”, 001\}.

A relation value is stored in a tabular file format physically (e.g., spss, csv, and sas). But this physical file is not what we refer to as physical description. A physical file is what can be downloaded and that can be in different file formats. In our study, the
**physical description** of a data set refers to a higher-level representation of the physical file – the heading and the body of a relation value, regardless of the order of attributes and tuples, or the medium of storage.

Figure 1: Our proposed framework for the representation of data sets using ontology

**Conceptual description level**

Conceptually, a data set is a collection of variable-value pairs related to a set of objects of study (often referred to in social science research as subjects, participants and respondents). The objects of study can be at different levels of granularity (e.g., individual persons, a group of people, an individual organization, or a group of organizations). The variable-value pairs represent concepts in a particular domain. We call it the **conceptual description** – the representation of the variable-value pairs as domain concepts for recording the characteristics of the study objects.

The separation of data set representation into the conceptual and physical perspective is to support data set reuse. The **conceptual description** represents information in a way that relates to users’ information needs. We assume that users searching and browsing in a data repository would be overwhelmed if presented immediately with tabular data sets. They are more interested in finding information about the objects of study (represented in the data set), and the variables in the data set that are related to social science research concepts. The **physical description** is needed for statistical analysis.
and data mining.

Thus, we propose that in the data repository, users should firstly be shown the data set’s conceptual description; the physical description is provided only if users request it. To link the conceptual description to the physical description, we defined a rdf: Property hasPhysicalRepresentation, as shown in Figure 1.

**Ontology level**

To represent the semantics of data sets, the conceptual description needs to be linked to an ontology. Both variables and variable values need to be mapped to concepts in the ontology (e.g. Gender: male|female). Concepts of social science variables and their value are collected from the sampled questionnaire, and are organized into an ontology with two classes – Class: Concept_Variable and Class: Concept_VariableValue. Both of them have subclasses and:

- Class: Concept_Variable have subclasses such as Class: Sex.
- Class: Concept_VariableValue has subclasses such as Class: Male and Class: Female.

There are relationships between these subclasses, as shown in the attached framework. We defined a rdf: Property hasConceptValue to link subclasses of Class: Concept_Variable and those of Class: Concept_VariableValue.

For a particular data set, its variables are instances of the ontology Class: Concept_Variable, and its variable values are instances of the Class: Concept_VariableValue. In this way, the semantics of a data set is assigned by linking its conceptual description to the ontology.

**Discussion**

The first half of this section is the discussion of the characteristics of our framework by comparing it with the Data Documentation Initiative (DDI) standard. DDI is a metadata standard developed for comprehensively describing and documenting social science data sets, and it is not designed to support social scientists reusing data sets. Our work address this issue by using the knowledge representation technique – ontology.

In the second half, we introduce a concept called valueType and discuss the reason and benefit of using it in the representation of social science data sets to support data reuse.

**Comparison with Data Documentation Initiative**

Data Documentation Initiative (DDI) is a standard for the documentation and management of research data created in the social science community. It is developed and maintained by the DDI Alliance, composed of an international group of data
specialists\textsuperscript{1}.

The reason for comparing our framework with DDI is that DDI is a well-known standard adopted by many data repositories for the description of data sets. Its development took into consideration 25 other standards (\textit{e.g.}, Dublin Core, Generic Statistical Information Model), and endeavoured to map and incorporate DDI to them (Thomas \textit{et al.} 2014).

Its most recent version (DDI Version 3.2) was published in May 2014. The standard takes into consideration the whole research data lifecycle: from a data set first being deposited into a repository, then modified, thus creating different versions, and finally being moved to other repositories. It provides metadata elements for describing data set related objects such as questions, variables and values in data sets.

The similarity between our framework and the DDI standard is that they both propose to describe data sets according to two facets; physical and conceptual. However, DDI is designed to support data acquisition, discovery and long-term preservation (Vardigan 2008), which is from an archival and documentation perspective. It is not designed to support the reuse of data sets by social scientists. Our framework thus aims to address this issue.

When reusing a data set, the first and foremost need of social scientists is to quickly understand what is inside it. However, this is not an easy task due to the issue of synonymy and homonymy of terms (\textit{i.e.}, different researchers may differently name variables that refer to the same or similar concepts. Alternatively, variables with the same name may refer to different concepts). To address this issue, our framework proposed to add an \textit{ontology level} as the top layer linking to the \textit{conceptual description level}.

This is missing in the DDI standard. Ontology assigns meaning to terms in the Conceptual Description Level (\textit{i.e.} variable name and value name), and links variables within and across data sets, thereby supporting unambiguous understanding of the concept of variables, as well as the discovery of variables of similar concepts.

Ontologies have been increasingly used in e-learning systems and digital library systems to support retrieval of resources. He \textit{et al.} (2010)’s study has shown that systems supporting ontologies outperformed those with only metadata descriptions of infrastructure in the retrieval of user-desired resources. The types of resources in these systems are usually text-based, such as journal articles or course materials.

User behaviour differs in relation to using text-based materials and data sets. Users can read the content of a text-document to assess its relevance, but are unable to assess the reusability of a data set by looking at its numerical content. The semantics of a data file is encoded in its direct expressions (\textit{e.g.}, numbers), and need to be decoded.

\textsuperscript{1} https://www.ddialliance.org/about/about-the-alliance
Thus, linking a data set to an ontology not only requires linking a main theme of the data file to the ontology, but requires linking expressions of the theme to the ontology as well. We think the expressions of quantitative social science data sets should include columns, rows, and values in the cell. They need to be represented and linked to an ontology. This is reflected in our framework (Figure 1).

In addition, we have developed an ontology modelling demographic information derived from social science research data sets, and the result is reported in a peer-reviewed conference paper (Sun and Khoo 2017).

**ValueType**

In the process of analyzing our sample (i.e., data sets and related questionnaires), we found a pattern in demographic and socio-economic variables across social science data sets: the specific composition of a set of answer choices for a question vary from one data set to another, depending on data collection country, time and research topic. For example, as shown in Table 1, we found these two sets of answer choices repeatedly occur in questionnaires in Singapore and the UK, respectively.

<table>
<thead>
<tr>
<th><strong>Singapore Internet Survey</strong></th>
<th><strong>General Household Survey 2000 (from UK Data Archive Survey Question Banks)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>1Chinese</td>
<td>1White</td>
</tr>
<tr>
<td>2Malay</td>
<td>2Black - Caribbean</td>
</tr>
<tr>
<td>3Indian</td>
<td>3Black - African</td>
</tr>
<tr>
<td>4Others</td>
<td>4Black - Other Black groups</td>
</tr>
<tr>
<td></td>
<td>5Indian</td>
</tr>
<tr>
<td></td>
<td>6Pakistani</td>
</tr>
<tr>
<td></td>
<td>7Bangladeshi</td>
</tr>
<tr>
<td></td>
<td>8Chinese</td>
</tr>
<tr>
<td></td>
<td>9None of these</td>
</tr>
</tbody>
</table>

As introduced in our framework, the Ontology Level comprises a class of variable concepts and another class of value concepts. For each variable concept, it can be linked to many value concepts. The pattern implies that these value concepts can be grouped into different subsets based on different contexts. Value concepts in different subsets may also overlap.

We designate these subsets of value concepts as valueType. A valueType is defined as a datatype property in the Web Ontology Language. The technical definition of valueType is presented in our paper (Sun and Khoo 2017) and will not be discussed here.

The standard to select a subset of value concepts and define them as a valueType is that this subset is common to a particular country or context (e.g., time), and has been
reused all the time in different data sets. For a particular questionnaire, if the values of a variable seem to be close to one of the valueType (e.g., a subset of the values in the valueType is used), we just use this valueType. After all, our aim is not to create valueTypes for each of the questionnaires, but to create unique valueTypes that can possibly show certain patterns of social science quantitative data sets.

The benefit of defining valueType is to select a set of concept values that do not overlap and mean something in a particular context. When named appropriately, it can help social scientists to make quick decisions when reusing data sets. For example, a valueType named U.S. Race 1 will let users infer that this subset of values make sense in the U.S., and is unlikely to be used in a Portuguese questionnaire.

Conclusions

This paper presents our knowledge organization framework for organizing and representing social science research data sets. The principle purpose of our framework is to support social scientists in reusing data sets. Our work complements the DDI standards by addressing this user perspective which is often overlooked.

References


The Genesis of Documentation in Brazil: Manoel Cícero Peregrino da Silva and Paul Otlet (1900-1924)

Abstract
At the end of the 19th century, the Belgians Paul Otlet and Henri La Fontaine created the International Institute of Bibliography, which, as well as providing a great source of universal information, pioneered new ways of processing documentary collections, thus ushering in design documentation. Between 1900 and 1924, the Brazilian National Library was headed by Manoel Cicero Peregrino da Silva, who brought about several changes to the institution, including the inauguration of a new building, reorganization of its collection, the creation of new services and, to some extent, the adoption of Otlet and La Fontaine’s ideas. This article aims to take stock of the contribution made by Manoel Cicero Peregrino da Silva and determine to what extent his actions contributed to documentation in Brazil and its engagement in the ideals of Paul Otlet and La Fontaine. This is a qualitative and quantitative study, of a historical-documentary nature, based on documentation from different Brazilian archives and from the Mundaneum archive in Belgium. It concludes that the role of Peregrino da Silva as director of the National Library made the institution a center where contemporary intellectual discussions converged. Moreover, several of Paul Otlet and Henri La Fontaine’s ideas were embraced by the National Library and it was the first institution to adopt documentalist practices in the treatment of its collection and the development of its products and services.

1. Introduction
In 1934, Paul Otlet published the “Traité de documentation le livre sur le livre, théorie et pratique”. Considered internationally to be the founding treatise in documentation studies, this publication was the culmination of ideals that the Belgian jurist had been formulating since the end of the 19th century. It was the result of reflections made when Otlet worked at the International Institute of Bibliography (IIB), as well as the ideals behind the construction of the Universal Bibliographic Repertoire (UBR) and his contact with institutions around the globe (López Yepez 1995).

The Institute’s aim, as a scientific association, was to develop an inventory, with systematic classification and description of the products of human spirit, that is, documents – understood here in their broadest sense. It thus sought to adopt a uniform and international classification system in order to assist the serious efforts of those wishing to publish, collect, analyze or consult books or any other products of human spirit, whether in the form of publications or whatever other possible means of documentation, (Institut International de Bibliographie 1895, 12-14).

Therefore, the creation of the Universal Bibliographic Repertory was in line with the missions of the institution, and was set up under guidelines that would allow it to be consulted and updated quickly, serving as a grand guide for those in search of information produced by people from all corners of the world, regardless of its origin or the format in which it was presented (Otlet and La Fontaine 1895, 16-17).
In 1911, the National Library (BN) of Brazil, under the directorship of Manoel Cícero Peregrino da Silva, embraced the universal ideal that lay behind construction of the UBR, by commissioning 600,000 index cards. However, the ideals of Otlet and Fontaine adopted by the National were not merely limited to the acquisition of index cards and Peregrino da Silva’s vision of the Belgian project was much greater than one may have thought.

From this perspective, we can point to the creation of the Bibliography and Documentation Service in correspondence with the International Bibliography Institute – responsible for the purchase of index cards; the centralization of exchange between national institutions and foreign institutions through the International Exchange Service; the approval of the National Deposit Law; the implementation of a series of conferences seeking to discuss matters of national interest; the reform of the institution’s regulations; and, finally, the inauguration of a new building, specially designed to house the National Library. We can also point out the adoption of documentalist ideals in the reorganization of its collection, as indicated by Peregrino da Silva in some of his reports and official correspondence.

This paper, which results from doctoral research in Information Science undertaken at the University of Brasília, makes an inventory of Peregrino da Silva’s actions and seeks to understand to what extent these actions contributed to the origins of documentation in Brazil and its engagement in the ideals of Paul Otlet and La Fontaine. Using a historical-documentary method, of a qualitative nature, the study is based on documentation in various Brazilian archives and in the Mundaneum archives in Belgium. Throughout this article, we aim to throw light on Otlet and La Fontaine’s ideals and how they influenced the origins of documentation in Brazil at the beginning of the 20th century, in the National Library under the intellectual leadership of Manoel Cícero Peregrino da Silva.

2. Documentation according to Manoel Cícero Peregrino da Silva

Manoel Cícero Peregrino da Silva was born on September 7th, 1866, in Recife (Pernambuco, Brazil), and died on October 3rd, 1956 in Rio de Janeiro. Graduated in Law, he acted in several fields during more than 50 years of professional activity, although, his most prominent role was as director of the Brazilian National Library, which he headed for 24 years, between 1900 and 1924 (Bittencourt, 1955).

Thus, we can identify the turning point that led to the complete reformulation of the National Library as being the “Project of the Regulation for the National Library of Rio de Janeiro”, elaborated in 1902 by Peregrino da Silva, which aimed to reorganize the services offered by the institution, as well as its spaces. A second landmark moment was the laying of the cornerstone of the institution's new building in 1905, which represented the first concrete step towards the library’s total reconstruction by which it
would undertake to better serve its users as well as accommodate its collection. This event initiated Peregrino da Silva’s planning of how to best occupy and use the new space, and, above all, how it could become the reference for all other libraries in the country, reflecting the modern ideals of reform underway in the then federal capital. A third important step was Peregrino da Silva's trip to Europe and the United States, with a view to seeing how librarianship and librarians were being developed there in the modern age.

A fourth notable marker can be identified in the inauguration of the new building in 1910, followed in 1911 by a crucial fifth milestone, namely the launch of the Library's new regulation, which, as promised in 1902, totally reformulated the institution and its services, seeking to honor that “magnificent palace” (Silva 1911). This scenario thus provided fertile ground for the adoption of new ideas and ideals, as Peregrino da Silva applied the thinking of Paul Otlet and Henri La Fontaine.

In fact, Peregrino da Silva had also improved and approved the Law of Legal Deposit established by Decree No. 1825, of December 20th, 1907 (Brasil 1907), requiring that typographies from all over the country send the institution a copy of each work that they published. What was most significant about this decree was the fact that it was the first time that a law of this nature was extended throughout Brazil, obligation to comply with such a law having previously been restricted only to the publishers of the then Federal District (Rio de Janeiro) (Fonseca 1973). This regulatory improvement was most likely a direct consequence of the Library's mission to safeguard national memory, as well as a response to another institutional goal: to make the Library honor its reputation as the greatest repository of knowledge in the country, and to offer its consultants a wider and up-to-date collection.

Another reason for the establishment of the Law of Legal Deposit is related to the fact that Brazil was one of the first signatories of the Berne Convention in 1886 to protect the rights of authors about their literary and artistic works.

In addition, adherence to the Berne Convention enabled the broader implementation of another service in the National Library, the International Exchange, because, in the reformulation put into practice by Peregrino da Silva, it would be the BN responsible for all the exchange of Brazilian institution’s with foreign counterparts, which means that the Library would become the center of distribution and reception of the most varied types of works, a fact that the Convention encouraged, now that the works were protected by an international agreement.

Another outstanding contribution was the Information Service, which once again characterized BN as a point of convergence for information, a stopping point (Costa, 2003) in the search for what one wanted.

The same regulation established the Bibliography and Documentation Service, which aimed to be the Brazilian base of the UBR, acting as one of its collaborators and
using the information listed there. Of course, the initiative to create the Bibliography and Documentation Service dates back to 1902, when in the “Project of the Regulation for the National Library of Rio de Janeiro”, Peregrino da Silva proposed the creation of an independent institution, attached to the Library.

As the driving force behind the plan to move the National Library to its new building, Peregrino da Silva wrote to the Minister of Justice and Internal Affairs, seven years after creating the prototype for the new regulation, to report on the events of the International Conference on Bibliography Documentation, held in Brussels. It was then that the National Library planned its move to a new space, while its techniques for treatment of the collection and the regulation of its documentation were revised. The director’s pronouncement to the minister, declaring his responsibility for the management and organization of all the production of human intelligence (Silva April 19th, 1909), clearly indicated the actions that he was subsequently to put into practice, and it thus is very probable that the desire for modernization, coupled with the institutionalization of the institution, prompted Peregrino da Silva to be part of the global integration project proposed by Otlet and La Fontaine.

The mission of the project, according to the then director, was to make all information available in an organized way, according to universal precepts and available to all people around the world, regardless of where such materials or persons were. As a partner in the initiative, the Library would no doubt consolidate itself as an information center, seeking to meet the wishes of its users, as well as the government's intentions to show that the country had modern institutions that left nothing to be desired by its European counterparts.

Thus, in the “New Regulation” of the National Library, the creation of the Bibliography and Documentation Service, by which it was the responsibility of the organization “[...] according to the system of decimal classification and by means of records, of the Brazilian bibliographic repertoire as a contribution to the universal bibliographic repertoire”, as well as “[...] the printing of these sheets to be displayed for sale or exchanged for foreign repertoire cards” and “[...] the organization of the collective catalog of Brazilian libraries” (Brasil 1911).

However, by our calculations, of the 600,000 index cards ordered for the IIB in 1911 (Rayward 1975; Juvêncio 2014), a little less than 330,000 reached the National Library, that is, 55% of the total number, which does not detract from the BN's commitment to the ideals of IIB.

It is undeniable, then, how the ideas and ideals spread by Otlet, La Fontaine and IIB had a strong influence on BN, a fact that can be explained by Peregrino da Silva's inspiration in the “modern”, as well as the fact that the institution was founded in an age that encouraged openness to new ideas. Therefore, with the IIB proposals as goals, the wide-ranging services implemented in the BN appear to finally legitimize the
contact that was being established between the two institutions. It also confirms Brazilian participation in the internationalist ideal of Otlet and La Fontaine. The greatest exponent of this interaction was the Bibliography and Documentation Service, together with the construction of an information network in Brazil, particularly through the initiative of creating collective catalogs and UBR records.

The Belgian partners’ aspiration to organize, treat and make all the fruits of human spirit available for consultation was, indeed, embraced by Peregrino da Silva, in order for information contained in the most varied types of support to be used for the development of education and research, as well as for promoting scholarship, something he always promoted.

We can thus readily perceive that the ideals of the universal Otletian project were made explicit and completely adhered to by the director of the BN. Acclaim was given to the idea of forming a network for the exchange and sharing of information, present in the restructuring process of the institution, with the objective of organizing documents of all shapes and sizes, and covering all their typologies. Its products – catalogs, bibliographies, index cards etc. – were to be constructed in accordance with an international standard in order to standardize descriptions, classifications and other representations. In fact, the great ideal behind these actions was to contribute to the organization of the UBR and the Universal Encyclopaedia.

Finally, Peregrino da Silva, envisioned that the project would only continue with the support of various sectors of society, from governments, publishers, researchers, working with the widest range of global institutions, as well as the construction and maintenance of collaborative networks and international agreements. The then director of the BN thus evidenced all his affinity with the ideals of Paul Otlet, Henri La Fontaine and IIB; after all, what he constructed was like a mirror of the initiatives the Belgians were carrying out.

Supported by a theoretical and legal framework allowing him to make changes that, in his view, were necessary for the BN, Peregrino da Silva put into practice a revolution in the way contemporaries viewed and conceived the National Library, or the "Palace" as he referred to the installations that the institution came to occupy.

Thus, at the time, Peregrino da Silva developed the BN into one of the largest ever projects for the organization and dissemination of information. The reforms he directed aimed at establishing BN among the most modern institutions of his age, for it to play a lead role nationally and internationally.

3. Final considerations

The ideal of creating a world-wide source of information encompassing all human knowledge was the basis of the philosophical conception of international cooperation advocated by Paul Otlet and Henri La Fontaine. In seeking to build a vast universal
catalog of human knowledge, the two jurists called on countries and organizations to work towards the goal of unifying all knowledge produced and recorded: to build the Universal Bibliographic Repertory.

The UBR, by itself, was already a daring proposal, since seeking to represent all the knowledge produced or to be produced by mankind was clearly not a simple task. One of the proposals of the Belgian partners was thus initially to standardize standards, adopting similar methods in all regions of the world, and to promote the use of the Dewey Decimal Classification (CDD) adapted for this purpose, the CDU.

In Brazil, Manoel Cícero Peregrino da Silva was the driving intellectual force behind the great transformation that the National Library underwent at the beginning of the 20th century. As we have seen, inspired essentially by the ideals of Paul Otlet and Henri La Fontaine, he promoted a series of changes, undertaken between 1900 and 1924, in the methods of organization and diffusion of information in the institution.

At that time, the BN was already the largest Brazilian library, but suffered from neglect and shortage of space, with its collection having been disregarded by the state for years. Although it was a symbol, its role was secondary and even unknown. Upon becoming a director, Peregrino da Silva adopted contemporary modern ideals and placed the institution at the service of the State, seeking to make it fulfill its mission of representing the country with the largest repository of the fruits of human spirit in Latin America, thus applying Otlet’s ideas to the maximum.

Moreover, institutions such as the Federal Senate, the Chamber of Deputies and the Royal Portuguese Reading Office were also to adopt the ideals of documentation, if not in their entirety, at least in part, when they began to organize their catalogs according to the guidelines of the IIB, and sought to disseminate them broadly, in alignment with the principles of the UBR.

However, the discontinuity of projects in our country is a striking feature that remains until the present day, with actions being in a constant state of deconstruction, remodeling or finalization at the whim of the latest intentions and personalities. As Fonseca (1957) confirms, the Bibliographical and Documentation Service did not escape this rule, being discontinued shortly after Peregrino da Silva left charge of the administration of the National Library and with him the integral interaction with the ideals of Paul Otlet and Henri La Fontaine.

In fact, the two world wars were crucial for the IIB project not to go ahead. Its collection had been damaged several times over the course of the two combats, and it perished also due to lack of resources and governmental disinterest, especially, as Rayward (1975) observes, after the appointment of Switzerland as neutral territory, with which Belgium competed. Subsequently, internationalist projects located in Belgian territory lost prestige despite the attempts of jurists.
However, this situation did not lead to Brazil’s complete abandonment of documentary ideals, since, despite having the IIB as a partner, Brazilian initiatives were firmly committed to creating a network of national catalogs, for example, and publishing the Brazilian Bibliography, as well, of course, as to maintaining constant exchange with diverse institutions.

"Refounded" 40 years later under the name of the Brazilian Institute of Bibliography and Documentation (IBBD), as Sambaquy (1956) tells us, the IBBD arose from the UNESCO initiative to promote the organization of information, especially scientific information, through the installation of institutes, a position which Otlet had promoted and encouraged since the late 1890s.

We can thus see that Manoel Cícero Peregrino da Silva, despite being an almost forgotten figure in the eyes of researchers on Information Science, Documentation and Library Science in Brazil, contributed greatly to developing these areas in the country. A product of his age, Peregrino da Silva was an intellectual who served the national interest, a man of visionary ideas who sought to promote the organization and dissemination of the BN collection, as well as adopting techniques that would be remembered again only 40 years after the founding of IBBD. This fact is demonstrated through the various documents recovered throughout our research. Guided by clues and traces left by other researchers into the origins of documentation in Brazil, we were able to unveil the influence of Peregrino da Silva and the key role he played in the origins of information science in Brazil – specifically, his contribution to documentation. In fact, we must pay attention to the significant addition to research provided through access to these still unpublished documentary sources, and we should be determined to use them to elucidate or at least shed new light on characters whose role cannot be denied in our understanding of how information science first unfolded in Brazil.

References
Brasil (1907). Decreto nº 1.825, de 20 de dezembro de 1907. Diário Oficial da União, 21 dec.


Deborah Lee, Lyn Robinson, David Bawden

Global Knowledge Organization, “super-facets” and music: universal music classification in the digital age

Abstract
This paper discusses the concept of universality in classification, by examining universal classification within a single subject area (music). It examines the idea of a “universal classification scheme” using music classification discourse, and analyses the concept of “super-facets” through three example sets of music facets. The paper finds three dimensions of universality within special classification schemes: coverage, mechanism and consumption. Furthermore, the “super-facet” is found to be a key part of universality, although its conception of universality is different from the universal classification scheme.

Discussions about universality in classification are an important part of recent knowledge organization (KO) discourse. For example, a call for a new universal classification by Dahlberg (2017) and a panel at the 2016 ASIST conference about global/local knowledge organization (Adler et al. 2016), are just a few examples showing the importance of these subjects to contemporary KO. This paper aims to look at a particular type of being universal: the need for, and development of central and universal classifications for a particular subject. In the case of this paper, music will be used as an example. As well as exploring the idea of a “universal classification” for a single subject (music), this paper will also deconstruct the idea of universality even further, by examining the idea of “super-facets”.

The paper starts with a brief review of universality within KO literature. The idea of a universal classification scheme for music is explored, and three dimensions of universality are posited. The idea of super-facets is then introduced, and three systems of super-facets for music are mined for information about how super-facets function. Finally, the relationship between universal classification schemes for music and super-facets is explored. So, while this paper focuses on what by some definitions of universal is impossible, “the universal special classification”, by doing so, further information about what exactly it means to be “universal” is laid out.

1. Literature review

Universality and its related concepts are important topics in KO discourse. KO literature suggests that a “universal classification scheme” is often taken to mean one which covers the whole universe of knowledge – for example, Dahlberg’s (2017) call for a universal classification scheme. In this respect, a universal classification scheme can be considered a particular type of general classification scheme, using the common division of classification schemes into general and special schemes. Furthermore, Szostak (2010) suggests that there is an important debate within KO which sees the
desire for a super, universal classification scheme, especially valuable to those working in interdisciplinary research, as diametrically opposed to the idea of domain classification – an opposition, it should be added, that Szostak (2010) attempts to breach in his paper.

However, theorists suggest that the term “universality” has a number of different meanings (Szostak, 2014; Satija and Martínez-Ávila, 2017); indeed, exactly what is “universal” even within a so-called universal scheme is not always explicit. In particular, Szostak (2014, p. 469), seeks to clarify what we mean by “universality” and outlines the difference between universality as the possibility of being able to “see the world in the same way” and being defined by “seeing the world in the same way”. (This idea links to discussions of universality in aesthetics, and whether universal aesthetic judgement is about the object’s potential or everyone having the same opinions about that object (McMahon 2007, p. 30-31).) The relationship between interoperability and universality is another important part of discussions (Gnoli and Szostak 2014); especially important for this paper is their idea of “conceptual interoperability” (Gnoli and Szostak 2014), in other words universality as a system of knowledge (as opposed to a technical mechanism) designed to work across different classification schemes. Also of interest is the part of universality which deals with the idea of global versus local classification (Adler et al. 2016); again, although the discussions might focus on general classification schemes, the overall principles of global classification practices versus local ones still has currency within domain-specific classification schemes.

So, this review of literature has shown that while the idea of universality is often seen as “… simply a system that seeks exhaustive coverage rather the treatment of a limited domain” (Szostak and Gnoli 2014, p. [2]), there are also other possible conceptions of universality at play, such as the generic idea of the global, universality as conceptual interoperability, and “universal” representing a singular viewpoint.

2. Aspirations of a universal music classification

Examining music classification schemes yields an important phenomenon: the aspiration and attempted realisation of the “universal music classification scheme”. These could be seen as special-classification siblings to the general, universal classification schemes which dominated the late 19th century and much of the 20th century. The ideal of being “universal” can also be seen as part of a wider movement, fitting into Gnoli and Szostak’s (2014, p. [1]) account of the popularity of the term “universal” in the 20th century.

This section is going to put the universal classification for music by Iván Pethes at its centre. A flexible classification system of music and literature on music (shortened to Flexible for this paper) was first published in 1967, and created partly as a discussion document for IAML and the International Federation of Documentation. Pethes (1967, p. 1) describes his scheme as being a starting point for “a universal classification system.
of music”. Hence, the aspiration for *Flexible* as universal music classification scheme is clear. However, what does Pethes intend by the idea of “universal”? There are a number of places where Pethes describes universality or global ideas, and these are valuable insights into this elusive concept of universality.

First, Pethes directly mentions the idea of universality in conjunction with what we might consider as traditional ideas of universality: the global, general classification scheme. Pethes (1967, p. 2) opens his introduction by mentioning “… the great universal classification systems of the 19th and 20th centuries, prepared by Harris, Cutter, Dewey, Brown, Bliss and Ranganathan …” and how these have the aim of “universality” (Pethes 1967, p. 2). So, we can infer that when Pethes uses the term “universal” in conjunction with his own special scheme (*i.e.* not traditionally “universal”), he is aware of the ideas of universality perpetuated by schemes such as DDC and Bliss.

Second, Pethes (1967, p. 2) lays out various criteria of what should be included in an “internationally acceptable classification” system. There could be different ways of interpreting the term “internationally acceptable” in this context; however, we would argue that while he does not use the term “universal” at this specific juncture, we can assume that Pethes was equating his criteria with the “solution” to the problems of music classification – in other words, his creation of a universal classification scheme for music. Some of these criteria directly delineate as aspects related to universality, and are worthy of discussion.

Criterion 1 (Pethes 1967, p. 2) says that an internationally acceptable scheme should cover the whole field of music. So by this, we imply musicology, performance, analysis and so on. However, this could also mean music of all cultures and types. So, according to Pethes, to be a universal scheme, you have to cover all knowledge *in the appropriate domain*. Note how this could be perceived as a domain-specific version of the exhaustive, general scheme which usually is labelled as “universal”. Criterion 5, states that “it [the classification scheme]” should lend itself equally well for development by library, documentation and mechanical sorting” (Pethes, 1967, p. 2). So, the universal scheme is one which could be used equally for any type of music information collection or task. Consequently, this criterion is about collections of knowledge. Criterion 6 follows a similar theme: collection size. Pethes (1967, p. 2) states that an internationally acceptable classification scheme should be applicable to collections of any size.

Criterion 7 states that a scheme needs “… an international terminology to overcome language difficulties” (Pethes, 1967, p. 2). While the word “terminology” could have a number of meanings (for example, names of instruments, genres, and so on), comments by Pethes about the failure of other schemes for not using international symbols, suggest that Pethes is here talking about the use of international *notation*, and is likely advocating the use of numbers and symbols over letters. The idea of internationality being expressed through notation is a fascinating interpretation of what it means to be
universal. Furthermore, this link between decimal notation and universality is not new; for example, Hornbostel and Sachs (1992) borrow the decimal notation of DDC for their 1914 classification of musical instruments, for reasons including the notation’s perceived universality. Notation-as-universality also acts in a different dimension to the other criteria discussed so far, as it concerns the mechanism of a classification scheme.

Flexible is not the only writer to consider what shape a universal classification for music might take. For example, Chailley (1988) also writes about the need for a universal classification for music, with a different focus from Flexible. Chailley (1988, p. 244) wanted universal coverage in terms of the formats of musical materials. So, we can add another aspect to Pethes’ list: format of music materials as a prerequisite of universality in a classification scheme.

Finally, there is another interpretation of “universal” which is not stated explicitly but implied in Pethes’ writings: the universal usage of a scheme. We can consider this in terms of the reception-based idea of “consumption” (Lee 2015), which looks at how often and where a classification scheme is used. So, while Flexible was designed to inspire international usage, there is little documentary evidence of its actual usage either contemporarily with its publication or the years since. Surely one part of being a universal and “internationally acceptable” classification scheme (such as Flexible), is for it to be used widely and internationally? This asks a challenging question about universality: can a scheme be universal if its usage is not universal? Flexible demonstrates how universality in design and intention are not equivalent to universal usage.

We can put together all the criteria together into a model of universality, and this model conceives of three dimensions of universality: coverage, mechanism and consumption – see Figure 1. Individual aspects are included within these dimensions, such as notation (in the mechanism dimension) and formats (in the coverage dimension). Furthermore, the aspects within coverage could be loosely divided into those which act at individual resource level (such as knowledge covered or format) and those which act at collection level (such as type of collection or collection size).

3. Super-facets for music

The search for universality in music classification is not limited to the creation of universal music classification schemes: specific individuals and groups have produced what they believe to be the facets of music, or at least the facets of Western, notated art music. We could call these facets “super-facets”. Three particular systems will be briefly described: a project by the IAML Sub-commission for Classification in the 1970s to produce a comprehensive set of facets for music (IAML facets); a set of facets produced Redfern in the 1980s as part of his textbook on music classification (Redfern facets); a series of super-facets devised by Elliker in the 1990s to analyse music classification systems (Redfern facets). In each case, we are more interested in the
universal motivations and what we can glean by this novel concept of super-facets, than what each system says about music. (More details about any of these systems of super-facets, a comparison of the systems, and facets of music classification more generally can be found in Lee (2017)).

The IAML facets were published in brief in *Fontes Artis Musicae* (Dorfmüller 1975). From this document we have a skeletal idea of where these came from, albeit lacking the details of how every facet was born. The five facets (Dorfmüller 1975, p. 48) are assumed to cover musical works and works about music. These facets appear to have been created to bring together existing thinking about music, and to solve standardisation issues within music classification. Interestingly, Dorfmüller (1975, p. 48-49) notes how some facets were easy to agree upon such as medium and time period, whereas others such as purpose/occasion/effect/intension required discussion. So, we could ask is a super-facet more or less universal if it is easily agreed upon by its authors?

The Redfern facets were designed for another reason: they appear as the inevitable result of facet analysis in a textbook about music classification (Redfern 1978). Redfern (1978) gives a total of 13 facets, some of which apply to music and works about music, while others are only for works about music. This is an interesting variation on the idea of universality across different formats of musical materials. While Redfern (1978) narrates the creation of his facets as inductive reasoning from the analysis of 26 examples of music resources, it is assumed that he purposefully selected these examples in order to “uncover” the specific list of 13 facets. The small sample size and manner of presenting the facets suggests that Redfern believed that there is a “single truth” of facets of music which would be revealed through facet analysis of *any* collection of musical knowledge. The Redfern facets imply that universal facets of music *exist*, and only await discovery, in contrast to the IAML facets which are derived from discussions about the multitudinous possibilities of music’s facets.

The Elliker facets have a different purpose. Elliker (1994) devised a set of seven elements for music (not works about music) in order to analyse 24 music classification schemes. He created these facets through inductive means, by analysing existing music classification schemes and music classification literature. Although not designated as facets by Elliker, these seven elements are used like facets by Elliker, and so will be considered as such. The aim of the Elliker facets was to develop a unified set of facets – he calls this a “metataxonomy” (Elliker 1994, p. 1271) – in order to gain the common terminology and structural units needed to analyse 24 classification schemes of music. In other words, the Elliker facets were designed for the purposes of universality as interoperability – more specifically, to borrow Gnoli and Szostak’s (2014) term, “conceptual interoperability”. In this sense, they are truly super-facets, as they specifically represent the commonality between facets found across multiple classification schemes. Intriguingly, the Elliker facets explicitly and deliberately utilize the IAML and Redfern systems: he merges the two existing systems (Elliker 1994, p.
1.270), as well as adding a facet of his own (Elliker 1994, p. 1.271), arguably creating “super-super-facets”.

Considering these three particular systems of music super-facets is useful for exploring the general idea of super-facets. To start, like a universal classification scheme for music, the presence of multiple systems of super-facets shows how there is no single set of universal facets. The different usages and intentions for the three sets of super-facets illuminate different approaches to universality. Redfern (1978) offers up facets of music, not super-facets; however, his assumption that these facets are the facets of music, is what (arguably) makes them super-facets. Both the IAML facets and Elliker facets are created for universality-as-interoperability; however, while the IAML facets appear to be created for a general ideal of standardisation, the Elliker facets have a prosaic and defined interoperability task awaiting their derivation (namely, the analysis of 24 music classification schemes). Finally, while these three systems of super-facets were created for bibliographic classification schemes in the 20th century, the idea of super-facets also has value in a digital environment; for instance, Downie’s (2003) facets of music have arguably been adopted as a sort of unofficial set of super-facets by the music information retrieval community, while Madalli, Balaji and Sarangi (2015)’s selection and utilisation of facets for their faceted ontology of music, in some ways echoes the creation of the Redfern facets (although not in the decided facets).

4. The connections between super-facets and universality

This exploration of universal classification schemes and super-facets for music, also questions the connection between universality and super-facets. For example, Flexible, a universal classification scheme for music, uses facets selected especially for Flexible rather than a system of super-facets. Similarly, the three examples of super-facets for music do not lead directly to a universal, or indeed any, classification scheme. So, we can see that super-facets do not appear to be components of a universal classification scheme, and that super-facets exist outside of the framework of universal classification schemes. Actually, the super-facets and universal classification schemes discussed in this paper work at different levels: for example, Flexible is a single classification scheme, whereas the Elliker facets are devised from multiple classification schemes and represent the consolidation of classificatory ideas from multiple schemes. Using the above music examples as a guide, it can be posited that a universal classification scheme attempts to establish universality by being a single, universal classification system, whereas a system of super-facets works amongst different classification schemes to establish universality across multiple schemes. So, super-facets might be universal, but are not part of the concept of a universal classification scheme.

Furthermore, we also need to consider where super-facets fit within the three dimensions of universality articulated above. The three examples of super-facets could be seen as a combination of the coverage and mechanism dimensions of universality;
however, rather than being a feature of a single scheme, super-facets exist as the shared space between the coverage and mechanism dimensions across *multiple* classification schemes. Thus, it is shown that super-facets and universal classification schemes may be different phenomena, but are both parts of the same dimensions of universality.

5. Future work and concluding thoughts

This paper has shown the value in considering universality in classification schemes within a specific domain. However, as this paper only discussed a small selection of schemes and super-facets within one particular domain, it would be interesting to see how and if they could be applied to other domains. Furthermore, it would be useful to see how the dimensions of universality articulated in this paper and the idea of super-facets, could be fitted into discussions about universality within general classification schemes. This paper asks, beyond anything else, what does “universal” really mean, using the environment of a specific domain to analyse unexplored parts of this concept. The answer, in part, complements existing literature about universal classification: even within a single domain, the concept of universality in classification is a significant concept, but its meaning and scope is anything *but* universal.

Figure 1: Three dimensions of universality in a classification scheme
References


Lilium Rajan

Historical ambiguity: a lens for approaching outdated terms

Abstract
Knowledge organization systems attempt to find a place for every resource, yet assorted leftovers seem inevitable. Focusing on the aspects that render something difficult to describe, rather than on the end result of their description, the author previously offered a taxonomy of types of ambiguity (Rajan, 2017). Among these types of ambiguity are resources that display historical ambiguity due to their use of terms that have changed in meaning over time. Here the author considers three cases of term change documented by scholars around eugenics, Asian American identity, and neurodiversity. The concept of historical ambiguity is proposed as that which collocative integrity seeks to address. However, while collocative integrity imagines some concept stability at specific moments in time, historical ambiguity also considers the fluidity of concepts in a particular moment. This fluidity may be particularly present in cases of human identity. The challenges that human identity presents to knowledge organization may be seen as related to its position as a thing-oriented rather than an experience-oriented discipline.

1. Introduction
The ideals of knowledge organization imagine a universe in which everything finds a place and can thus be readily discovered and obtained. We seek to create an order in which everything is properly accounted for. Practitioners know well the challenges of the misfits, the errant, defiant or unknowns. Bowker and Star’s Sorting Things Out highlighted the multiple effects that these supposedly neutral systems have, including the ways in which they exclude (1999). In recent decades, many scholars in knowledge organization and beyond have illuminated the existence of myriad others. In knowledge organization, others may be evident via explicit labeling as such but more often they disappear from view through alternative placements. We have come to understand a range of consequences of othering, including misrepresentation, invisibility, exploitation and reification. Following on the work of scholars who have documented these consequences, I have begun to ask what characteristics render some things other. In a previous paper I suggested framing these characteristics as ambiguity (Rajan 2017). By considering these characteristics rather than the consequences of representation, I hoped that we might find new solutions toward future design. In that taxonomy, I proposed four broad types of ambiguity, each of which might have a different solution. In this paper, I revisit what I termed historical ambiguity to see how the concept overlaps with or extends work that addresses a similar work, especially subject ontogeny.

To do this comparison, I examine three papers that engage changes in term use over time: the ontogeny of eugenics in the Dewey Decimal Classification (DDC) (Tennis 2012); the history of Asian American terms in the DDC (Higgins 2016); and a proposed
taxonomy of terms related to neurodiversity (Zolyomi and Tennis 2017). These cases each illustrate different aspects of the challenges of term change over time, though they are far from exhaustive in documentation of these issues. Nonetheless, they allow us to consider the ways in which the concept of historical ambiguity (or ambiguity more broadly) may serve those engaged in the work of categorizing and designing systems of representation. Finally, I will briefly consider two ways of engaging with the concept of ambiguity and ask how these approaches highlight specific challenges in the way that knowledge organization considers what is possible and necessary.

2. Background

In classification, there should be a place for everything. What then of those things that inevitably defy ready categorization? There are many ways that these challenges are handled, from being scuttled into a junk drawer category such as “other” or being forced into another category that may or may not have a clear relationship to the resource in question. In addition to being confusing or useless, these misrepresentations can perpetuate and enforce power over marginalized identities, as has been illustrated by Olson (2002), Adler (2009) and Fox (2016b), among others.

There are several ways resources can end up as others. In the first case, one might know what something is, but the system has no place for it. In other cases, the issue is not merely a missing term. Instead, something remains opaque or confusing. I have characterized this quality as ambiguity. In looking at how ambiguous others are handled by classification systems, it is helpful to consider the many ways in which a thing can be ambiguous.

This focus is an extension of questions brought forth by Star and Bowker in their discussion of residuality, a term they use to describe that which remains unclassifiable due to issues of temporality and personal experience or perspective, such as fluctuating physical pain (2007). It is not possible to pin these things down in representation, rendering these realities as formally non-existent. This work has been taken up by other scholars in relation to resource description, including Feinberg, Carter and Bullard’s work on video metadata (2014).

Taking otherness and residuality to refer to the same struggle in representation, there are several ways in which resources with these characteristics may be handled. Such a resource may be up-posted, in which case its particulars are lost. It may be discarded. In other cases, it may be falsely named, in which case it may disappear while counting as a representation of something else entirely. It may be declared other or some variant, usually nested within a classification. I have characterized this problem as one of ambiguity. The terms other and residuality both emphasize where things are left (or left out), while ambiguity shifts the attention to the moment before resources are described and asks us to imagine what has confused the description process.
In this paper, I will extend my previous work that proposed a taxonomy of four types of ambiguity (Rajan 2017), one of which, historical ambiguity, I will consider further here. Each type is united primarily by potential solutions. This taxonomy is system agnostic, with examples gathered from literature and conversations with practitioners in multiple disciplines. The first type of ambiguity is perhaps the simplest: multiplicity. This is used to describe resources that have a large number of equally relevant aspects or concepts, which challenges most knowledge organization systems. Nonetheless, a solution would be to assign as many terms as possible to arrive at an adequate representation. The next type of ambiguity I called emergence, the formation or arrival of a concept. This is particularly common in scientific communities. The solution would be something like patience. What I called privacy-related ambiguity is seen in cases where humans are engaged in their own classification, or description of resources from their community. This kind of ambiguity arises largely from lack of trust of the system of categorization. The solution is building trust between the community and those who seek to describe them. Finally, I introduced conditional ambiguity, which describes resources whose identity is so context specific that they may not be accurately represented without some form or narrative description. While otherness may be related to all four types of ambiguity, residuality corresponds most closely to this final type.

Within conditional ambiguity, I offered several examples, including historical ambiguity. This term was intended to describe resources that use terms considered out of date. For this project, I considered how historical ambiguity maps with other ideas about challenging representations.

To do this work, I considered three papers published in the past decade that look at changes in terms over time. These papers do not exhaustively map the field. In the following sections, the term “resource” will be used to refer to that which is classified or categorized. This term is not always a perfect fit, particularly in reference to people.

3. Historical ambiguity: three cases

3.1. Tennis

We begin with the subject ontogeny of eugenics by Tennis (2012). Since it was published, the approach that Tennis takes has been replicated by numerous others in analysis of historical changes to classification systems. In this paper, Tennis traces the wide range of subject locations for works on eugenics in the DDC from 1911 to the present. After being banished from the biological sciences in the 1950s, the subject does not have a tidy new home. Instead, it is characterized by lumping and splitting from edition to edition. Further, Tennis notes that at times resources are classified in defiance of the most recent editions of DDC, sometimes conforming instead to earlier editions, perhaps in order to maintain coherence in collections.

While there are many concerns with the representation and misrepresentation of a
subject with so fraught a history as eugenics, Tennis notes that from the perspective of knowledge organization, this creates particular challenges to the goal of *collocative integrity*, or the appropriate and comprehensive gathering of related resources in a collection (2012, 2013).

If we apply the lens of historical ambiguity to this case, we might note that there are two particularly challenging aspects. First, the subject may become illegible to users, who may find resources on eugenics in places that are contradictory or confusing, without being able to understand or reconcile these contrasting perspectives. Secondly, as Tennis highlights, the professional who seeks to classify a resource that indicates a biological scientific perspective on eugenics is prevented from doing so in accordance with the current edition of the DDC. Thus, the subject loses salience.

### 3.2. Higgins

Many scholars in recent years have put an ontogenetic lens to terms related to human identities and traced the consequences of these classifications. Representation of Asian American subjects, as explored by Higgins (2016) is particularly interesting for the expansiveness of the communities which this term covers. While Higgins examines editions of the DDC for placements of Asian American subjects from 1876-1996, she notes that the term “Asian American” didn’t emerge as a collective identity until the late 1960s, and it only makes an explicit appearance in that form in the 1996 edition. Therefore, the ontogenetic question in this case is how a multitude of identities that may be now described as Asian American were represented in the DDC in this period. While many identity terms appear and disappear over time, most are inadequate at distinguishing Asian subjects from Asian American ones. Further, in those cases where the context clearly indicates an Asian population in the United States, it tends to be negatively formed. Higgins notes that the emergence of terms stems from literary warrant, so if multiple resources address the notion of cheap Chinese labor, for example, it may make an appearance as a subject term. While bibliographic classification systems aim to be able to represent all possible knowledge, they do so as need arises, rendering incomplete the range of possible subjects.

This criticism can be easily enough answered by the pragmatic purposes of a classification system, not to mention the impossibility of imagining all possible subjects divorced of literary warrant. However, it highlights historical ambiguity in pointing to the absences of terms that the user might seek. Further, the ontogeny that Higgins traces suggests that users (and possibly information professionals) must have knowledge of terms that are no longer in use, and which may be considered unjust, in order to discover older relevant resources.
3.3. Zolyomi and Tennis

The last case that will be considered is not centered on bibliographic classification, though it directly addresses term change over time. Zolyomi and Tennis perform a domain analysis of neurodiversity, with a particular focus on autism (2017). From this work, they offer a taxonomy of terms related to experiences of neurodiversity, both social and medical. The case of autism shares characteristics with our previous examples. On one hand, autism is a diagnosis and therefore linked with the scientific community. Terms related to autism are found in multiple classifications, including the Diagnostic and Statistical Manual of Mental Disorders (DSM) and the International Classification of Diseases. In this sense, the existence or reality of autism may be seen as related to the case of eugenics. On the other had, a diagnosis (or lack thereof) of autism has an impact on human identity. Thus, if the scientific view of autism changes, it has a direct effect on those whose identity is related to that diagnosis. Indeed, this can been seen in the case of Asperger’s, a diagnosis related to autism until it was eliminated in the DSM-V in 2013. Many people found the diagnosis and the identity useful for understanding and advocating for themselves or their communities. Under the guidelines of the DSM-V, there would not be any “new” individuals with Asperger’s, even if they fit the criteria previously associated with the diagnosis. The taxonomy offered by Zolyomi and Tennis retains this term and includes many concepts that are related to autism but are not formally accounted for in current classification systems. This taxonomy includes concepts about lived experiences of neurodiversity, including an in-between state of identity in relation to autism, which is given the term neurocurious. If we consider this state in relation to historical ambiguity, agnostic of any particular knowledge organization system, we may appreciate the challenge of representing identities that are subject to multiple temporal perspectives. That is, terms may change in their denotation and connotation, as eugenics has. They may also change in actual terminology, as illustrated by Higgins through terms of Asian American identity. Finally, they may change in and of themselves, and not necessarily in a linear fashion. Identity can be contextual and contingent. If our systems are unable to note context, it is possible that resources will become misrepresented or invisible.

3.4. Summary

In each of these cases, we have seen examples of terms that are not easily understood over time, and thus ambiguous. I characterize them specifically as historically ambiguous because they highlight the challenges of consistently describing certain concepts and related resources over time. These examples bring our attention to two very different sources of subject terms. In the cases of eugenics and neurodiversity, the terms of use have emerged from scientific and medical literature. Their use claims the reality of the perspectives they represent. Tennis has traced the move eugenics made in the DDC, from its origins in biological sciences to its current banishment from those
classes. Zolyomi and Tennis bring our attention to the medical origins of terms related to neurodiversity and note that as the official medical classifications change, our sense of what is real, what can count, is troubled. Is eugenics science? Does Asperger’s exist?

Bibliographic classification systems are generally inclined to support literary warrant and thus to continue use of terms that are frequently used in the literature. Therefore, in spite of the elimination of Asperger’s in the DSM-V in 2013, we still find Asperger’s syndrome as a topic (a narrower term of both Autism spectrum disorders and syndromes) in the Library of Congress Subject Headings. At what point do these classifications intersect, or does the scientific discourse become preferred over the social discourse?

The cases of Asian American subjects and neurodiversity illustrate the ontogenetic challenges of representing human identity. Both of these terms contain a multitude of communities with identities that may or may not align with terms used by classification systems. Deciphering the perimeters of these terms requires much greater engagement with these communities than mere consultation with the latest scientific literature.

In all of the cases examined here, the primary challenge they present is one of collocative integrity (Tennis 2012, 2013). If we assign terms to resources in different ways over time, we will fail to bring them together in ways that benefit the user. While Tennis highlights this problem, it is imagined to be potentially solvable through phoenixing, or reclassifying entire subjects and all related collections. Alternatively, one might consider legible notes in the record that would alert the user to ontogenetic changes and direct them to related relevant resources. However, especially in cases of human identity, phoenixing can be just as harmful as leaving offensive or inaccurate terms in place, as has been explored by Furner (2007). Ontogenetic annotations may miss the point, too. In these cases, ambiguity may be more perpetual and less easily resolved. Even if we accept the concept of literary warrant, how do we want to reconcile terms that have been used violently with terms that are liberatory?

This paper does not seek to answer this question but notes that the conversation is well captured in recent work by scholars in multiple disciplines. Adler has proposed the concept of taxonomic reparations as a means to bring greater justice in representation of historically marginalized groups (2017). This work aligns with the concept of cognitive justice, especially in relation to indigenous knowledge organization (Mouliason, Sandy, and Bossaller, 2017). The challenges of representing intersectionality have recently been explored by Fox (2016a). The problem of intersectionality is not necessarily historical, yet it has always existed and adds another challenge to our analysis of historical term change, as alluded to by Higgins. An applied approach to reconciling multiple representation requirements can be seen in the double sets of metadata used by Glass in relation to First Nations objects housed in museum collections around the world (2015). These are but a few conversations related to the
problems of historical ambiguity. All of these approaches recognize the problem of
collocative integrity, but they do not necessarily privilege it. The question remains who
this form of integrity serves, and how. This exploration of historical ambiguity suggests
that some cases may be more readily assessed in terms of this metric than others.

My interest in considering ambiguity is in capturing the residual. In this case, it
might be said that the method of ontogeny maps historical ambiguity. However, the
work of ontogeny is always backward-facing and two-dimensional. It is determined by
the systems that have organized the universe and recorded how resources fit into it.
These systems always have bias, as well documented by numerous scholars, and so
ontogeny work still can only highlight what has been done, not what could have been
at any point in time. As such, it may suggest a greater fixity in concepts of the past,
even as they change in time. In many cases, especially of human identity, the semantic
values of terms can multiply swiftly without losing their primary meaning, as in cases
where slurs are reappropriated by marginalized communities.

More than simply being the landscape that ontogeny maps, historical ambiguity
should include the unmapped territory, the terms that might have been, or that are so
contextual that they hold for only a moment.

Further questions
In creating this taxonomy, I acknowledged multiple sources of ambiguity: that of the
indexer, that of the user, and, perhaps, that of the resource itself. In attempting to parse
these questions, there are two ways of considering ambiguity. First, we can see
ambiguity as a thing, or as a property of a thing. This is the assumption embedded in
much of my taxonomy, derived from the thing-oriented discipline of bibliographic
classification. In this field, we look to resources to derive our descriptions. We try to
remain objective and descriptive in our use of concepts. We want to be able to create
rules that can be used by multiple professionals and be consistently understood by most
or all users. There is inevitable compromise in this, but it is generally acceptable and
domain-specific (Hjørland 2009). However, human identities are not a domain. People
are not resources in which one can find precise attributes that allow us to make the right
call. Furner grappled with this in his analysis of identity (2009).

This tension suggests that in some cases, ambiguity can be understood as a thing or
a characteristic of thing, but more often it is an experience. We encounter, as
professionals or end users, and wonder to ourselves what precisely something is. In
some cases we can point to the sources of our confusion, but the confusion is ours. The
world is not entirely concrete, yet knowledge organization wishes to render it so. What
are we failing to capture? Gorichanaz (2016) has characterized this split as one between
veritas, which emphasizes factuality, and aethesia, or uncovering, and suggests that
knowledge organization is preoccupied with veritas at the expense of the much more
challenging alethia. This may map well to the differences between ambiguity as thing versus experience.

It remains to be seen how useful ambiguity is as a lens for considering the challenging outliers in knowledge organization and classification. Taxonomizing the unclassified may be a perverse exercise, however suited it is to the practices of the disciplines it engages. It is possible that everything that has here been characterized as ambiguous in fact is only ambiguous to someone at a certain point in time. If we can pinpoint the precise context of both the resource and viewer, perhaps all ambiguity can be resolved. The question then is: is it possible for anything to be ambiguous to itself? Can ambiguity be an inherent trait? We might look to conversations about the relationships between thoughts and words, to the champions of representation who claim that we cannot be what we cannot see. It seems possible that there are situations in which an identity cannot easily be positively asserted, where a total absence of vocabulary renders representation impossible but through the act of opting out of available terms. The cases where this holds may be too infrequent to be of concern to knowledge organization professionals. However, it may be helpful for us to be able to triangulate our unsettled confusions by specifying a vocabulary for ambiguity.

References


Lynne Bowker

How Knowledge Organization helped to shape the emerging field of Terminology in Canada

Abstract
As the field of terminology began to take shape in Canada in the 1970s and early 1980s, it became clear that traditional linguistics methods were not sufficient to support this new field of activity. Following an analysis of five seminal Canadian works on terminology published during this period, we illustrate that knowledge organization had a significant influence on the development of terminology methodology by contributing ideas relating to conceptual maps, semantic relations and thematically structured presentations.

Introduction
Terminology is a relatively new field of linguistic activity described by Sager (1990, 2) as “the study of and the field of activity concerned with the collection, description, processing and presentation of terms, i.e. lexical items belonging to specialised areas of usage of one or more languages.” Terminology is often understood in contrast to lexicography. While lexicographers compile dictionaries containing general language words, terminologists prepare glossaries of terms from specialized fields of knowledge. Discussions about the need for terminology as a field of activity first surfaced in the 1930s, when an Austrian engineer named Eugen Wüster (1931) emphasized the need for clarity and precision in technical communication. However, it was not until the 1970s that the field of terminology as we know it today really began to take shape. While terminology is firmly rooted in linguistics, there is widespread recognition that it is an interdisciplinary field. For instance, in the introduction to the first volume of the international scientific journal Terminology, the editors observe that “Terminology … is not based on linguistic principles alone, but is itself essentially of a multidisciplinary nature” (Loening and Sonneveld, 1994, 3). They, along with other terminology researchers, such as Sager (1990, 3), Picht and Draskau (1985, 22), and Cabré (1999, 25), refer to contributions from disciplines that include cognitive science, computer science, and information science, among others. However, beyond this type of general acknowledgment, we know of no detailed analysis of specific contributions made by information science – and more specifically by knowledge organization – to the formation of this new field of activity.

This paper seeks to explore some ways in which knowledge organization contributed to the formative years of terminology, focusing in particular on ways in which ideas and practices from knowledge organization inspired the development of an initial methodology for conducting terminology work in Canada in the 1970s and early 1980s. To the best of our knowledge, no prior research has focused explicitly on identifying
the ways in which methods and resources developed and used in knowledge organization were borrowed, adapted and integrated into the working practices and education of Canadian terminologists during the early years of this new field.

The paper is divided into four main sections. First, we describe the context in which terminology work developed in Canada. Next, we introduce the corpus of five seminal Canadian terminology publications that forms the basis for this study and that is investigated for evidence of contributions from knowledge organization to the emerging discipline of terminology. This is followed by a presentation and discussion of the themes identified in the corpus. Finally, we offer some concluding remarks.

1. Translation as a driving force for the emergence of Terminology in Canada

As Rondeau (1981, 38) notes, in the years following Wüster’s efforts to encourage clear and consistent technical communications, different groups began working in the area that came to be known as terminology. Consequently, different schools of thought emerged, including the German-Austrian school, the Soviet school, the Czechoslovakian school, and the (Canadian) Quebec school. It has been noted that, among these, the Quebec school distinguished itself by its close ties to translation, as well as by its strong focus on methods and pedagogy (Rondeau 1981; Delisle 2008).

Although terminology work can be carried out in a monolingual context, in the officially bilingual country of Canada it is strongly associated with translation (Kerpan 1977, 46-47). By the 1960s, translators working for the Government of Canada’s Translation Bureau estimated that they spent more than one-third of their time researching specialized terminology (Dubuc 1972, 36). To increase translators’ productivity, it would be necessary to reduce the amount of time they devoted to terminology research. Therefore, in Canada, terminology began to emerge as a new field of activity that was auxiliary to translation and was carried out by terminologists.

Commenting on the state of terminology in Canada near the end of the 1960s, Delisle (2008, 124) notes that terminology was not taught, textbooks did not exist, terminological research methods were embryonic, computerized term banks had not been developed, and the tasks of terminologists had yet to be defined. Nonetheless, driven by the needs of the rapidly expanding translation industry, terminology training was clearly needed. Not surprisingly, therefore, early efforts at providing formal terminology training took place within translator education programs. Indeed, each of the 12 undergraduate translation programs that were established in Canada in the 1970s and 1980s included at least one terminology course (Delisle 2008, 273). Where there is teaching, there is also a need for teaching materials, and it was during the 1970s that the first serious efforts to develop and document a methodology for conducting terminology work in a Canadian context took place. In the period between 1973 and 1981, Canadian professors, translators and linguists produced five seminal publications on terminology. The product of careful reflection and experimentation, these five
ground-breaking contributions, some of which generated further editions, would prove to have staying power. The next terminology books written chiefly for a Canadian audience did not appear for 20 years (i.e., Pavel and Nolet 2001; L’Homme 2004).

2. General methodology and corpus description

Our approach to learning more about how knowledge organization influenced and inspired the early development of terminology in Canada has been to closely read the five Canadian publications on terminology that were produced between 1973 and 1981 to identify themes that reveal underlying contributions from knowledge organization. In this section, we will provide a general introduction to the five works in the corpus, before going on to discuss their knowledge organization-related content in more detail. In summary, these five seminal works from the Quebec school, published in French, served to document the development and evolution of a methodology for terminology work and served as teaching material to train Canada’s first generation of terminologists. The list below outlines key features of each volume.

1) Corbeil (1973) Guide de travail en terminologie (Guide to working in terminology): The first volume on terminology to be published in Canada, it was a modest effort intended to launch discussions on the development of a methodology for conducting terminology work. It was based largely on the professional tips and tricks gleaned through the experience of employees of Quebec’s Office de la Langue Française (OLF).

2) Auger and Rousseau (1978) Méthodologie de la recherche terminologique (Methodology of terminology research): A reworked version of the initial OLF volume, this work took into account the experience and knowledge gained by attempting to implement earlier versions of a terminological methodology. More widely distributed, it helped the Quebec school to gain international recognition.

3) Boutin-Quesnel et al. (1979) Vocabulaire systématique de la terminologie (Systematic vocabulary of terminology): A complement to the OLF’s methodology document, this structured glossary of terminology vocabulary was the first of its kind in the field and was intended to describe the metalanguage of terminologists. It had a unifying effect on Canadian terminologists and helped to establish terminology as a distinct field.


5) Rondeau (1981) Introduction à la terminologie (Introduction to terminology): The
first Canadian volume to break away from a purely practical perspective, this work nonetheless provided an overview of approaches to terminology, and helped to situate the work that was being carried out in Canada in a wider international context.

3. Findings and discussion

In this section, we will first present a general overview of the main steps that emerged as being important for conducting terminological research. Having this high-level understanding of how a terminologist works will help readers to better process the more specific discussion that follows, which considers how principles and practices from knowledge organization contributed to the development of a methodology for the newly emerging field of terminology in Canada.

3.1. Overview of the main steps in a Terminology research project

The goal of a terminology research project is to produce a glossary or collection of term records that contains a comprehensive coverage of the terms belonging to a specialized field (or subfield) of knowledge. A glossary typically contains a structured list of the preferred terms, accompanied by additional linguistic information such as a foreign language equivalent, related terms (e.g. synonyms, abbreviated forms, spelling variants), and part of speech. The term is also accompanied by a definition, as well as a context showing an example of the term in use. Sources are clearly documented. An alphabetical index accompanies the structured collection to facilitate look up.

To produce a glossary, Canadian terminologists working in the 1970s typically followed these main steps. First, terminologists did some background reading to familiarize themselves with the specialized subject field. Next they compiled a documentary corpus that they then examined to identify the terms that are proper to the subject field at hand. As they read the corpus, terminologists extracted not only the potential terms, but also additional information that could be used to help them map out the conceptual structure of the field and understand the relations between the concepts. Information that could be used to help define concepts or demonstrate examples of the term in use was also collected. The data were then analyzed to complete tasks such as identifying which term would be the preferred term (and which would be designated as synonyms), or defining the term. In Canada, where terminology was almost exclusively carried out in a bilingual fashion, all of these preceding steps were conducted independently in both English and French, and then an additional interlingual analysis was carried out in order to establish a conceptual match between the concepts in both languages and to confirm linguistic equivalence. Finally, the results would be compiled into a glossary or collection of term records.
3.2. Contribution of Knowledge Organization to Terminology

In this section, we present a number of themes that were identified in the five works described above, focusing specifically on those terminological activities that were inspired by work being carried out in knowledge organization.

Before there was an established methodology for carrying out terminology work in Canada, those charged with conducting terminological research experimented and did the best they could, drawing mainly on approaches used by 19th-century lexicographers (Delisle 2008, 42). However, lexicographers take a semasiological approach to their work, which means that they begin by identifying the lexical item and work towards establishing its definition. In other words, lexicographers ask the question “What does the word X mean?” It soon became clear that this approach had limitations in terminology, where it was important to understand the subject field as a whole rather than considering the terms in isolation. Indeed, a more appropriate question for terminologists seemed to be “What do you call X?”, which takes the concept as the starting point, rather than the term. Therefore, these early terminologists soon began looking beyond lexicography to draw inspiration from other disciplines too, including knowledge organization. An examination of the contents of the glossary of terminology prepared by Boutin-Quesnel et al. (1979) reveals a number of entries that have been borrowed or adapted from knowledge organization:

- analyse notionnelle (subject/concept analysis)
- classement systématique (subject order)
- index (index)
- notion (concept)
- relations internotions (semantic/conceptual relations)
- terme générique (broader term)
- terme spécifique (narrower term)
- terme privilégié (authorized term)
- terme rejeté (non-authorized term)
- vedette (heading/entry term)

Information about knowledge organization proved to be useful to terminologists in a number of different ways, including helping them to produce a conceptual map for the subject field under investigation, helping them to identify the conceptual relations needed to create definitions and establish interlingual equivalents, and helping them to present the results of their research in a structured format.

3.2.1. Subject field breakdown

For instance, Corbeil (1973, 28) advises that an early and important step in a terminology project involves delimiting the domain that will be the subject of the research. He indicates that it is not sufficient to simply give the name of the domain,
but rather, it is necessary to clearly specify the subdivisions or branches of the domain that will be taken into account, as well as those that will be excluded from the project. He suggests that a terminologist can facilitate this task by taking as a starting point an existing classification, such as the Universal Decimal Classification (UDC).

Auger and Rousseau (1978, 17) take things further, noting that for each subdomain, the terminologist should prepare a structured list of concepts. According to Auger and Rousseau, the elaboration of a concept system makes it easier for a terminologist to get a better overview of the subject field and to understand the relations between the concepts. Auger and Rousseau (1978, 20) likewise suggest consulting thesauri and subject classifications as models to help guide this process, a recommendation that is later echoed by Rondeau (1981, 72).

Dubuc (1978, 36) is the first from the Quebec school to use the word onomasiological to describe the concept-to-term direction of terminology research, and like Auger and Rousseau, he advocates strongly for the elaboration of a concept system that illustrates how the various concepts in a subject field are related to one another. Not only is the resulting conceptual map useful for providing the terminologist with a general overview of the subject field, it also provides vital information for later stages of the terminology research project, such as definition construction and the establishment of interlingual equivalence.

3.2.2. Establishing definitions and interlingual equivalents

After noting that definitions represent one of the most complex aspects of terminology work, Corbeil (1973, 26) explains that one very effective way to define a term is to refer to its broader term (i.e., a more generic concept) and to indicate how it differs from its coordinate concepts. This method is also advocated by Dubuc (1978, 98), and to do this, the terminologist must have a good understanding of the semantic relations in play. Indeed, Boutin-Quesnel et al. (1979, 27) define the term definition in a way that draws attention to the importance of conceptual relations for definition construction: Definition: ‘a statement that describes a concept and allows it to be differentiated from other concepts within a concept system’.

Meanwhile Dubuc (1978, 72) explains, and Rondeau (1981, 33) confirms, that the way to establish whether English and French terms are equivalent is to determine whether both refer to the same concept. This requirement comes out clearly in Boutin-Quesnel et al.’s (1979, 20) definition for Equivalent: ‘each of the terms of different languages that designate corresponding concepts’. Hence, once again, understanding the place of a concept within the concept system is critical.

3.2.3. Systematic presentation of entries

The emphasis on knowledge organization in terminology also extends to the way the contents of glossaries are organized. In contrast to lexicography, where the vast
majority of dictionaries use alphabetical ordering to present their entries, Corbeil (1973, 67) and Auger and Rousseau (1978, 46) encourage a systematic organization for terminology glossaries, identifying a number of benefits to this latter approach. For example, in the same way that the conceptual map could help the terminologist to gain a better understanding of the overall subject field, so too could a structured presentation help the glossary user to better understand how the different concepts are related to one another. Corbeil (1973, 68) recognizes that systematic ordering requires a more complex design and a greater effort on the part of the terminologist than does alphabetical ordering, but he argues strongly that it is more advantageous for the user. Nonetheless, when a terminological resource is ordered systematically, Corbeil recommends providing a corresponding alphabetical index to facilitate look up.

Boutin-Quesnel et al. (1979) put this recommendation into practice when producing the glossary of terminology. The glossary adopts a systematic ordering for its entries, which are divided into three main categories (terminology theory, methodology, and types of terminology collections), with further levels of subdivision. The systematic presentation is accompanied by an alphabetical index.

4. Concluding remarks

As pointed out by Delisle (2008, 169), Canadian linguists first began to reflect seriously on the theory and methodology of terminology in the 1970s. It was during this decade that Canada’s terminology community began to organize itself, to equip itself with conceptual tools and a metalanguage, to outline working methods, and to submit their ideas to the test of experience. Early efforts drew heavily on methods used in lexicography, but it soon became clear that these alone would not suffice and that budding terminologists would need to look elsewhere for additional inspiration.

One of the disciplines that terminologists turned to was knowledge organization. The first four terminological works to be published in Canada were highly practical in their orientation and contained very few references to existing literature. However, Rondeau (1981) incorporated more theoretical ideas as well. This latter volume includes a 16-page list of references, including a number drawn from the field of knowledge organization. Among these, for example, we find references to several early works by Ingetraut Dahlberg, a German professor of information science who founded the journal *International Classification* in 1974, the title of which was changed to *Knowledge Organization* in 1993, and which remains important in the field today.

Through a close examination of five seminal works published by members of the Quebec school between 1973 and 1981, we have demonstrated that knowledge organization significantly influenced the emerging field of terminology in Canada, and particularly the development of a methodology for carrying out terminology work. As emphasized by Rondeau (1981, 61), methodology has an extremely important place in terminology, and the strong focus on methods and on bilingual comparative
terminology are features that distinguished the early Quebec school from others.

While terminology’s linguistic and lexicographical roots remain evident, there can be no denying that this young field also owes much to the discipline of knowledge organization, which inspired many elements of the concept-oriented methodological framework needed to carry out terminological research effectively.

Acknowledgements

An expanded version of this paper investigating how terminology has been influenced by the field of information science more broadly is found in Bowker (2017).

References


Abstract

Metatheory, in general, is the analysis of theory and there are at least four types of metatheory: Metatheorizing in order to better understand the theory (Mu); metatheory as a prelude to the production of a new theory (Mp); metatheorizing to produce a perspective that overarches some part or all the theory (Mo) and; metatheorizing to evaluate a theory (Ma). This paper proposes the analysis of the influence of metatheoretical investigations, published in the journal Knowledge Organization, on the KO domain. The theoretical discourse of KO regarded to metatheory is shaped it seeks to support the understanding of its different approaches in the KO domain. The citations to metatheoretical investigations are collected from the databases: Web of Science, Scopus and Google Scholar. The intellectual influence from metatheoretical investigation to the citing papers is analyzed considering the possible errors regarded to citation analysis.

1. Introduction

Metatheory is a common research method in Information Science (Metcalfe 1957; Zhao 1991; Svenonius 1992; Vakkari and Kuokkanem 1997; Cronin 1998; Hjørland 1998; Vickery 1998; Olson and Schlegl 2001; Tennis 2005, 2008; Dousa 2010; Tennis 2015; Tognoli 2015; Araújo, Guimarães, and Tennis 2017). According to Tennis (2008) this research has in common the intention to provide a narrative about theory.

A significant part of Knowledge Organization (KO) research is regarded to the epistemological, methodological and theoretical issues. Hence, we consider metatheory, while not always explicitly acknowledged as a method, it is fundamental to understand the foundations, the development of research, and the influence from other domains on KO.

Araújo et al. (2017) present an analysis of five metatheoretical studies (Tennis (2008); Dousa (2010); Samuelsson (2010); Castanha e Grácio (2014) and, Tennis (2015)) through Ritzer’s metatheory perspective. As a next step, we seek to analyze the influence of metatheoretical investigations published in the journal Knowledge Organization on the KO domain, in this paper. This way, we seek to shape the theoretical discourse of KO regarded to metatheory and to support the understanding of its different approaches in the KO domain.

2. Metatheory

"Metatheory, in general, is the analysis of theory", (Tennis 2015, 214). Metatheory comes from sociology and it is meta-analytic work. We agree with Hjørland that “philosophical knowledge has been neglected n IS and the epistemological and metatheoretical views have seldom been formulated or analyzed”. That is why, metatheory should be applied more often to knowledge organization (KO) research, “to
raise its theoretical and philosophical level, the better to understand the limitations and possibilities of different approaches” (1998, 620).

“Metatheory should be specified into unit theories by placing them in concrete social settings. One can say that a concept in a metatheory consists of a range of variation of meanings fixed by the definition of the concept,” (Vakkari and Kuokkanem 1997, 453). Furthermore, we can state that metatheory is “the set of assumptions presupposed by any more or less formalized body of assertions” (Bullock 1988).

Metatheory “presents all the situations or states of affairs that can be expressed within the logical possibilities of the conceptual apparatus of a theory” (Vickery 1998, 453). Following this thought, Vickery points out that to achieve success through metatheoretical work, it is important to make connections between the presuppositions that already exist in the domain, to show the weaknesses, and to propose alternatives (1998, 458).

In this research, we employ Ritzer’s metatheory (1990, 1991a, 1991b, 1992, 2001). Ritzer (1990, 3) “believes that much metatheoretical work has been done under a wide range of other headings – sociology of sociology, sociology of science, sociology of knowledge, history of sociology, and most notably, as an integral part of sociological theory”.

Zhao (1991) examines recent developments in meta-study in sociology which involves metatheory, meta-method, and meta-data-analysis. The author states that metatheory is the study of extant theory and it is probably the most fully developed branch of meta-study in sociology.

Ritzer (1991a) presents three types of metatheoretical work: - Metatheorizing in order to better understand the theory (Mu). "Mu is concerned more specifically, with the study of theories, theorists, communities of theorists, as well as the larger intellectual and social contexts of theories and theorists". It also offers systematic methods of understanding, evaluating, criticizing and, improving extant theory. - Metatheory as a prelude to the production of a new theory (Mp). It entails the study of extant theory in order to produce new theory. - Metatheorizing to produce a perspective that overarches some part or all the theory (Mo). The study of theory is oriented to the goal of producing a perspective, one could say a metatheory, that overarches some part or all the sociological theory. Mo involves a systematic study of the theory, it derived from theory.

A fourth type of metatheory was presented by Colomy (1991, 269), the idea of metatheorizing to evaluate a theory (Ma). "Ma refers to metatheorizing oriented toward devising and applying explicit, universalistic criteria to adjudicate the competing claims issued by rival social scientific traditions".

There are some critiques to metatheoretical studies. For Ritzer (1991a) scholars conduct metatheory unconsciously, but it would be more productive if metatheorizing occurred in a more self-conscious way. Vickery (1998) and Hjørland (1998) agree that
many metatheorizing works remain at too general level.

We could confirm those critiques in our study on the types of metatheoretical studies in the KO domain. (Araújo, Guimarães, and Tennis 2017). And, because we aim to explore the use of metatheory in KO, we propose this research. We present an overview of the influence of metatheoretical research on the KO domain.

3. Influence of metatheory in Knowledge Organization

There are many reasons authors cite (Brooks 1986). The authors may or may not agree with the argument from the citation they are using. But, by citing a work, authors acknowledge it and give it credit. Points of view, statements and, perceptions are connected and the author adds his own thoughts and ideas through citation. By doing so, the researcher creates connections between different researches and authors.

We acknowledge that there are some limitations in the analysis of the intellectual influence of citations in a domain. We agree with Zucherman (1987, 330) that "citations can be used as approximate indicators of influence for aggregates of authors and papers […]". Self-citation, incompleteness, bias in favor of elite scientists, under-citation and over-citation of fundamental scientific work are some examples of errors that underestimate and misrepresent the intellectual influence of citations in a scientific work.

We collected the papers’ citations to the metatheoretical investigations published in the journal Knowledge Organization to analyze the influence of metatheoretical investigations on the KO domain. (Figure 1).

Figure 1: Citations to metatheoretical investigation in databases

<table>
<thead>
<tr>
<th>PAPERS</th>
<th>CITATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WoS</td>
</tr>
<tr>
<td>Castanha; Grácio</td>
<td>1</td>
</tr>
<tr>
<td>Dousa</td>
<td>2</td>
</tr>
<tr>
<td>Tennis (2008)</td>
<td>11</td>
</tr>
<tr>
<td>Tennis (2015)</td>
<td>2</td>
</tr>
<tr>
<td>Samuelsson</td>
<td>4</td>
</tr>
</tbody>
</table>

Source: Araújo; Tennis 2017

We analyze citations indexed in Web of Science, Scopus and, Google Scholar where we identify some kind of metatheoretical influence. Self-citations are also considered. Eventhough self-citation is not conventional in this kind of study (because the author cannot influence himself), we include self-citation here to analyze how authors are using their own work on the same subject (Figure 1).

The citing/cited network of Web of Science is presented in the Figure 2 and, it is followed by the discussion regarded to the network.
Castanha and Grácio (2014) have one citation in Web of Science from Hilário, Tognoli, and Grácio (2016). The authors apply a metatheoretical investigation to the theme scientific collaboration. We identify a self-citation here, since Grácio is a coauthor in the citing and cited research. The authors cite the concept of metatheory presented by Castanha and Grácio (2014) to support their argument on the importance of metatheoretical investigations in information science.

Tennis (2008) receives eleven citations in Web of Science and we find different approaches in citations to his study. Tognoli (2015) cites Tennis (2008) as one of the authors that studies metatheory in KO.

Observing the limitations of intellectual influence through metatheory, we can see some kind of influence from Castanha e Grácio (2014) to Hilário, Tognoli, and Grácio (2016) and from Tennis (2015) to Tognoli (2015).

The citing/cited network of Scopus is presented in the Figure 3 and, it is followed by the discussion regarded to the network.
Castanha and Grácio (2014) have two citation in Scopus and, one of them is the same indicated in WoS (Hilário, Tognoli, and Grácio 2016). Lian et al. (2016) cite Castanha and Grácio (2014) when they describe domain analysis, since the authors establish a relation between bibliometric studies, metatheory and domain analysis.

Tennis (2008) has sixteen citations in Scopus and ten of them are the same from Wos. While addressing future research, Melgar Estrada (2011) cites Tennis’ (2008) work.

The author explains that Tennis (2008) presented a classification of KO research to the KO community which was situated in a metatheoretical framework. Melgar Estrada (2011) not only considers that the framework proposed by Tennis (2008) would be suitable to provide directions for future research on that subject. She also uses the framework to suggest some possible research themes on Topic Maps and KO. We can see the influence of Tennis’s metatheoretical research on Melgar Estrada’s work (2011), since she based part of her work on the classification created by Tennis (2008) using metatheory.

It is interesting that in another study that cites Tennis’ work (2008), Guimarães et al. (2014) state that domain analysis started to be studies through metatheoretical research. There is another citation to Tennis in that paper but it seems to refer to Tennis’ work published in 2003 since they cite the two axes of domain for domain analysis.

Castanha and Grácio (2014) study the bibliometric contributions to metatheoretical and domain analysis studies. They reflect on Tennis’ thoughts (2008) to show the importance of metatheory combined with domain analysis studies in the KO domain. We can observe the citation to that paper as foundational to support their argument. Hilário et al. (2016) do the same, since they also develop a metatheoretical study. They cite Tennis (2008) as one of the metatheoretical studies in KO and they also state the importance of applying metatheory and domain analysis together.
Melgar Estrada (2011), Castanha and Grácio (2014) and, Hilário, Tognoli, and Grácio (2016) were influenced by Tennis (2008) metatheoretical investigation as a foundational work on the use of metatheory.

The citing/cited network of Google Scholar is presented in the Figure 4 and, it is followed by the discussion regarded to the network.

Figure 4: Citing/Cited network (Google Scholar)

Source: Araújo; Tennis 2018.

Castanha and Grácio (2014) receives only one citation in Google Scholar, from Lian et al. (2016) as described before. Hjørland (2014) states that faced analysis should be considered a rationalist position and his argument was criticized by some authors, including Tennis (2008). Hjørland (2014) discuss Tennis’ argument that Ranganathan theory is not a strict rationalist stance, but more of a pragmatic, if not neopragmatic epistemic stance and method. And, in that study we can see the influence of metatheoretical study on the KO research.

There are two papers with self-citations from Tennis and in one of them the author only indicates Tennis (2008) study in the references. On the other hand, Tennis (2012) cite his (2008) work four times. Tennis (2012) explores two weights of language and action that bear down on KO researchers.

First, he presents the concept of KO from the (2008) study. Second, he develops his argument based on the forms of epistemology, theory, and methodology in KO and, following that, he presents an explanation on epistemology and epistemic stances.

Lastly, he discusses epistemic stance, methodological action and theoretical perspective in KO. He argues that it is possible to reflect on the weight of language and action in the context of KOS even there are competing theories of theory, what we consider in this study metatheory.
4. Conclusion

We present three citation networks created based on the citations to the metatheoretical investigation published in the journal Knowledge Organization. By analyzing those networks, we get to explicit facts like: - Tennis’ study on metatheory is the most cited, followed by Samuelsson and, Dousa; - Four papers (Dousa 2010; Castanha and Gracio 2014; Hilário, Tognoli, and Grácio 2016; Tognoli 2015) approach metatheory as a subject of the study.

Twelve of the papers analyzed in this study were published in the journal Knowledge Organization. Which demonstrates the strong influence of the journal in the domain. Most citations are related to the concept and discussion of key concepts in the KO domain. Some of them develop an argument and critique about the epistemological influence in KO research (Dousa 2010; Hjorland 2014; Szostak 2011).

Castanha, Grácio (2014) and Tennis (2008) studies were cited by other two metatheoretical studies, (Hilário, Tognoli, and Grácio 2016; Tognoli 2015). They use the citation to Castanha, Grácio (2014) and Tennis (2008) to support their argument and discussion on metatheory.

We can see the intellectual influence from Castanha, Grácio (2014) and Tennis (2008) on those new metatheoretical investigation. And, they have a specific influence on the others, probably, because they explicitly approach metatheory on the study.

The metatheoretical investigation does not influence only other metatheoretical investigation but support conceptual work like Melgar Estrada’s study (2011). The author uses the classification, created by Tennis (2008) through metatheory, to suggest some possible research themes on the subject of her research.

The self-citations analyzed in this research were basically used by the author to support the development of their research and continue the thoughts they presented in the original research.

Finally, we see some kind of intellectual influence through the citation analysis in the papers: Melgar Estrada 2011; Tennis 2012; Hjorland 2014; Castanha and Gracio 2014; Tognoli 2015; Hilário, Tognoli, and Grácio 2016; Hilário et al. 2017). We carefully consider in this research the possible errors related to intellectual influence through citation analysis.
References


Information and argument structures in Sociology research abstracts

Abstract
In a research abstract, information and arguments have to be organized and presented in a way that helps the reader to understand the research study and to be persuaded of the research arguments. This study analyzes the information and argument structures of 33 sociology research abstracts, focusing on 13 abstracts reporting investigative studies. The types and sequences of argument claims were analyzed, as well as the conceptual structure of each argument claim. Two semantic frames – the research-relation frame and the comparison frame – were applied to analyze the types of information provided, as well as how the semantic frames were progressively filled along the argument chain in the abstracts.

1. Introduction
Research papers typically contain an abstract with an organization structure that is a variation of the well-known IMRaD format: introduction, method, results and discussion/conclusion/contribution (Sollaci & Pereira 2004). Content-wise, an abstract provides an overview of the research study, including the research objectives and often an outline of the research method and results. A good abstract will also make an argument for why the research objective is worth investigating (e.g. by pointing out a research gap), and will justify the validity of the research results and the significance/contribution of the study. The information content and arguments have to be expressed textually in a way that helps the reader understand the information provided, and be persuaded of both the explicit and implicit arguments put forward.

This study analyzes the information and argument structure of sociology research abstracts. An argument can be divided into the argument claim (e.g. that a research objective is well-founded and worth investigating) and supporting argument (e.g. a research gap). The argument structure of a text (e.g. abstract) is the pattern of argument claims and supporting arguments represented in the text. This study first identifies the types of argument claims found in sociology research abstracts, and then analyzes the sequences of argument claims (referred to as argument chains) to identify argument chain patterns.

Information structure in this paper refers to the conceptual structure of argument claims and supporting arguments. An argument contains several pieces of information that are linked together. The argument claim often has a complex structure: for example, a research result may specify that a concept A causes another concept B in a particular context. The information structure can be represented as patterns of concepts and relations, represented graphically as a network of nodes (concepts) and edges (relations). This study makes use of the idea of semantic frames from frame semantic
theory (Fillmore 1976) to represent common information structures or patterns. The basic assumption of frame semantic theory is that the meaning of a word is embedded in and understood within a set of related concepts representing the essential knowledge related to a particular type of event or situation. The set of related concepts can be represented as a semantic frame specifying the expected relations/roles/slots and the types of concepts that can fill those roles.

In this study, we developed a few semantic frames to represent common situations represented in a research text, the types of information that are expected in the text, and the roles the pieces of information play in the situation. This paper focuses on two important and prevalent semantic frames found in research papers:

- the research-relation frame represents a relation (e.g. cause-effect relation) between two concepts, as well as the types of information that are relevant to the relationship, including the modality (i.e. true, false, probable, etc.), the size of the relation, and the evidence for the relationship. The Research-relation frame is often found in those sentences outlining the objectives and results of research.
- the comparison frame represents comparisons between two or more concepts, as well as the result of the comparison. It is often used to compare attributes of concepts and research outcomes in sentences outlining research objectives and results.

2. Literature review

The concept of argument in academic writing has been defined in different ways. It has been characterized as a main idea referred to as a “claim” or “thesis statement” supported by evidence (The Writing Center, University of North Carolina at Chapel Hill., n.d.). Some arguments are obvious and easily recognized by the reader because they “make a direct claim based on or drawn from evidence” (Lunsford & Ruszkiewicz, 2016, p. 5). Other arguments are indirect and need to be inferred from the surrounding text, knowledge of the text genre (in this case, research abstract), and even the content domain (sociology research).

A commonly used argument model in the field of academic writing and genre studies is Toulmin’s (2003) model of argumentation, which indicates that the argument claim may have qualifiers and rebuttals, and the supporting argument includes the data or evidence, the warrant (that authorizes the inference step from the data to the claim), and backing (which provides support for the warrant).

Although Toulmin’s (2003) model has been used to analyze and evaluate the argument structure in various educational contexts, several scholars (e.g. Wingate 2012) have noted that it can only model micro-level argument instances instead of macro-level argument structures. Toulmin, Rieke and Janik (1984, p. 14) referred to
this macro-level argument as “the sequence of interlinked claims and reasons that … establish content and force of the position for which a particular speaker is arguing”. In the present study, we analyzed the sequence of argument claims in research abstracts at the macro-level, which we refer to as the argument structure.

We also analyzed the conceptual structure of each argument at the micro-level using the research-relation frame and the comparison frame to identify the types of information represented in each argument. Ou, Khoo and Goh (2007) have previously analyzed the conceptual structure of research objective sentences in sociology dissertation abstracts. The research-relation frame in this study is an extension of their variable-based framework. In addition, we analyzed how the instantiated research-relation frame evolves along the argument chain—that is, how each argument along the chain adds new pieces of information to the research-relation frame, or modifies the existing information.

3. Methodology

The data for this study were abstracts taken from 33 journal articles published in eight sociology journals with high impact factor listed in InCites Journal Citation Reports. Our preliminary analysis indicated that different types of research have different argument and information structures. In our sample of 33 articles, we identified six types of research: historical analysis, descriptive study, development study, evaluation study, identification study, and investigative study. Due to space constraints, this paper reports only our analysis of 13 abstracts that describe investigative studies, which make up 39% of the sample. We define an investigative study as research that aims to investigate a relation (i.e. cause-effect, association, or co-occurrence) between two concepts/entities, often by carrying out a survey or experiment.

Having previously developed three semantic frames in a preliminary study (Cheng, Khoo and Kathpalia, 2017), we herein describe and illustrate two of them, namely the research-relation and comparison frames, using an example abstract given in Table 1. This abstract (with the last two sentences left out) is part of a journal article reporting an investigative study to identify causal relations between two entities – embedded social media in hotels.
### Table 1: Example abstract and chain of argument claims found

<table>
<thead>
<tr>
<th>Sentences in the example abstract</th>
<th>Argument claims</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Support] Considering the strong [Research-relation] influence of social media on internet users,</td>
<td><strong>Argument 1:</strong> Claim – Topic Centrality</td>
</tr>
<tr>
<td>[Claim] it is important to understand its role for hotel businesses, particularly the online aspect of lodging operations.</td>
<td><strong>Information structure:</strong></td>
</tr>
<tr>
<td></td>
<td>[social media] –influence-&gt; [internet users].</td>
</tr>
<tr>
<td></td>
<td>[hotel websites] –attr-&gt; [social media].</td>
</tr>
<tr>
<td>[Claim] Although several social media studies have been done, very few studies have focused on travelers’ needs and the specific gratifications they seek when using embedded social media channels on hotel websites, and how those channels would [Research-relation] influence their purchasing behavior.</td>
<td><strong>Argument 2:</strong> Claim – Research gap</td>
</tr>
<tr>
<td></td>
<td><strong>Information structure:</strong></td>
</tr>
<tr>
<td></td>
<td>[Internet users: travelers] –attr-&gt; [gratifications]</td>
</tr>
<tr>
<td></td>
<td>–attr-&gt; [purchasing behavior].</td>
</tr>
<tr>
<td>[Claim] The main purpose of the current research is to examine the effectiveness of embedded social media channels on hotel websites and their [Research-relation] influence on traveler behavior.</td>
<td><strong>Argument 3:</strong> Claim – Objective/question</td>
</tr>
<tr>
<td></td>
<td><strong>Information structure:</strong></td>
</tr>
<tr>
<td></td>
<td>[affects] –cause-&gt; [social media] –attr-&gt; [hotel website]</td>
</tr>
<tr>
<td></td>
<td>–effect-&gt; [Internet users: travelers]</td>
</tr>
<tr>
<td>[Claim] Applying the uses and gratifications (U&amp;G) approach, we examined [Research-relation] relationships among traveler gratifications, satisfaction, and purchase intentions by [Comparison] comparing user experience with hotel websites that used embedded social media channels to those without embedded social media channels.</td>
<td><strong>Argument 4:</strong> Claim – Objective/question (narrower)</td>
</tr>
<tr>
<td></td>
<td><strong>Information structure:</strong></td>
</tr>
<tr>
<td></td>
<td>[travelers] –attr-&gt; [gratifications]</td>
</tr>
<tr>
<td></td>
<td>–attr-&gt; [satisfaction]</td>
</tr>
<tr>
<td></td>
<td>–attr-&gt; [purchase intentions]</td>
</tr>
<tr>
<td></td>
<td>[comparison] –cause-&gt; [hotel website] –attr-&gt; [social media]</td>
</tr>
<tr>
<td></td>
<td>–effect-&gt; [hotel website] –attr-&gt; [NO social media]</td>
</tr>
<tr>
<td>[Claim] The results indicated that travelers exposed to the hotel website with embedded social media channels had [Comparison] higher levels of perceived informativeness, perceived enjoyment, and perceived social interaction that directly [Research-relation] influenced traveler satisfaction.</td>
<td><strong>Argument 5:</strong> Claim – Result</td>
</tr>
<tr>
<td></td>
<td><strong>Information structure:</strong></td>
</tr>
<tr>
<td></td>
<td>–effect-&gt; [perceived informativeness, enjoyment, social interaction]</td>
</tr>
<tr>
<td></td>
<td>–size-&gt; [higher]</td>
</tr>
<tr>
<td></td>
<td>[perceived informativeness] –affects-&gt; [traveler satisfaction]</td>
</tr>
<tr>
<td>[Claim] In the context of embedded social media channels, the gratification factors, such as perceived enjoyment and perceived social interaction, directly [Research-relation] influenced traveler satisfaction and purchase intentions, and indirectly [Research-relation] influenced purchase intentions through traveler satisfaction.</td>
<td><strong>Argument 6:</strong> Claim – Result (narrower)</td>
</tr>
<tr>
<td></td>
<td><strong>Information structure:</strong></td>
</tr>
<tr>
<td></td>
<td>[perceived enjoyment, social interaction] –affects-&gt; [traveler satisfaction, purchase intentions]</td>
</tr>
<tr>
<td></td>
<td>[perceived enjoyment, social interaction] –affects-&gt; [purchase intentions]</td>
</tr>
</tbody>
</table>

websites and traveler satisfaction. The first column lists the sentences in the abstract, and the second column indicates the types of argument claims represented in the sentences.

Figure 1 illustrates an instantiated research-relation semantic frame, showing the information structure of a particular research-relation specified in the example abstract of Table 1. Most investigative studies seek to identify specific relations (e.g. cause-effect and association) between important research concepts or entities. We refer to all these as research relations. The research-relation concept is linked by the relations concept1 and concept2 to two related concepts. In Figure 1, concept1 has been specialized to a more specific relation, cause. Concept2 is specialized to effect. Each concept may have one or more attributes and aspects. In Figure 1, attributes of the research-relation concept include size, and a direct or indirect relation. The instantiated research-relation frame shown graphically in Figure 1 represents the information structure of the abstract in Table 1.

4. Result

4.1. Analysis of argument chains

In a research paper, many types of arguments are made. In a preliminary study (Cheng et al. 2017), we identified eleven types of argument claims: centrality (or importance) of the topic, general statement (generalization), research gap, research issue/problem, research idea/approach, objective/question, hypothesis, method, result, development of something new, and contribution. We also identified ten types of supporting arguments.

We analyzed the argument chains found in the abstracts. For the example abstract from Table 1, the chain of argument claims is listed in the second column, and summarized as follows:

Topic centrality \(\rightarrow\) Research issue/gap \(\rightarrow\) Objective/question – Objective/question (narrower) \(\rightarrow\) Result
– Result (narrower)

This represents a fairly complete argument chain. From our analysis, we derived the following comprehensive argument chain for research abstracts:

General statement \(\rightarrow\) Topic centrality \(\rightarrow\) Established knowledge (from previous studies) \(\rightarrow\) Research gap \(\rightarrow\) Objective/question \(\rightarrow\) Method \(\rightarrow\) Result \(\rightarrow\) Contribution/Recommendation

We shall refer to this as the full research argument chain. However, few abstracts use the complete argument pattern, most following a sub-sequence of the pattern. In the 13 abstracts in question, the objective/question and the result appear very frequently, in 12 (92%) and 11 abstracts (85%), respectively, while contribution is present about half of the time, in 6 abstracts (46%).
In 4 of the abstracts (31%), the objective/question is followed by one, or more, narrower objectives/questions. 5 abstracts (38%) show that a given result may also be followed by a narrower or more specific result. It is well-known that a research gap/problem/issue is an important argument and justification for carrying out research. We found that only 5 abstracts (38%) pointed out a research gap/problem/issue.

Overall, the abstracts follow the sequence of argument claims specified above, with one exception. One abstract asserts the centrality of the topic after the objective/question statement, probably to highlight the importance of the objective.

4.2. Analysis of information structure in the arguments

We analyzed the overall information structure of each abstract, by identifying the pieces of information in the text that fill different elements/roles in the research-relation and comparison frames, illustrated in Figure 1. Overall, cause-effect and association relations are found in almost all abstracts reporting investigative studies. Context information is provided in 69% of the abstracts, and modality in 46%. Furthermore, all
the research-relation frames are linked to one or more comparison relations, typically comparing attributes or aspects of a concept, and comparing alternative research-relations (representing different outcomes in the study).

We also analyzed how the research-relation and comparison frame elements are progressively specified along the argument chain. We focused our analysis on the following frequent argument progressions:

- Case 1: Established knowledge or Topic centrality → Research gap
- Case 2: Research gap or Topic centrality or Established knowledge → Objective/question
- Case 3: Objective/question → Result
- Case 4: Result → Contribution or Recommendation

For Cases 1 to 3, the Research-relation frame for the supporting argument is usually underspecified, and the argument claim will then add more detailed or specific information. Looking at Case 2 in more detail, a research gap statement typically indicates either an unknown element in a research-relation frame, or a general concept that needs to be specialized with a narrower concept. So the research objective statement that follows the research gap statement often provides more detailed information by:

1. Specializing the research-relation, for example from association in the research gap to cause-effect in the research objective statement.
2. Specifying the cause and/or effect concept in more detail, for example specifying a subclass or an attribute of the cause concept that is the actual causal factor.
3. Specifying two or more subclasses of the cause and/or effect concept for comparison.
4. Specifying additional types of information as constraints on the study.
5. Alternative causal concepts or factors

An exception is the case of established knowledge from previous studies → research gap or objective/question. The research gap and objective statement may seek to broaden the research-relation that was established in previous studies. Case 4 (Result → Contribution or Recommendation) also usually involves broadening or generalizing the research-relation information.

5. Conclusion

We have analyzed the argument chains in 13 sociology research abstracts reporting investigative research studies, and derived a comprehensive chain of argument claims. Most abstracts use a sub-sequence of this chain. We also identified the most common argument claims, and alternatives to the common argument claim sequences. We also analyzed the information structure with the help of two semantic frames (research-relation frame and comparison frame) that specify the types of information that are expected to be represented in the abstract. We identified the most common types of
information found, and how the semantic frame elements are progressively filled or modified from a supporting argument to the argument claim.

This study will be extended to other types of research studies (e.g. qualitative descriptive studies and evaluation studies), and carried out on the other sections of the research paper (e.g. Introduction and Literature Review). The results of this study can be applied to the teaching of academic report writing as well as inform research in natural language processing, especially in automatic discourse analysis and summarization of research papers, and in argumentation mining.

References


Wieslaw Babik

Information and knowledge ecology: a field for research in Knowledge Organization

Abstract
This article concentrates on information and knowledge ecology, considered to be one of the most recent fields of information research in Poland and abroad. I intend to present the subject matter, the research context, Polish and foreign researchers, current achievements, relationships towards other research fields, and then postulate on future directions of development. These topics will be discussed in the context of general trends in scientific development, as well as in the field of knowledge organization. The goal of my article is to show the main problem areas within the research field of information and knowledge ecology, which is currently considered to be one of the most promising directions in research development, including information science. This area shows clear signs of becoming an independent research field as indicated by the fact that some researchers have assigned it the status of a separate scientific discipline.

Introduction
We are living through times when there is a very clear opposition between the development of our civilisation and the requirement of information protection. Information ecology is considered, next to information architecture, to be one of the most recent specialties in the humanities, including information science. The field started to develop in response to the dominating pressure imposed by certain technological aspects of design and expansion of information systems in wide-area computer networks at the turn of the 21st century. Information ecology, being a complex of issues and actions associated with information protection against “information pollution,” is a remedy for such pollution threats (Babik 2008). Within that trend, the appearance of numerous studies on neuroergonomics can clearly be witnessed.

Although information and knowledge organization encompasses a wide range of issues concerning information and knowledge ordering to facilitate data search, it is not conducive to dealing with the excess of information sources combined with highly problematic information content value in an Internet environment where a huge quantity of dispersed, fragmented, and unordered information is available. Information explosion, especially in cyberspace, and, consequently, permanent increase in information resources, cause many difficulties in reaching information and knowledge. Information and knowledge are essential elements of man’s diverse and dynamically developing information environment, which is facilitated by the existence of many different systems of organizing information and knowledge. In the event of catastrophic information trash, it is imperative to struggle against the phenomenon of information overload. We are also faced with the problem of evaluating the credibility of
information available in network resources. It has become necessary to eliminate, or at least limit, the existing barriers in the use of information and knowledge, by creating appropriate methods and tools at the same time as adequately evaluating and organizing information quality. If information is supposed to be the substance for building and organizing knowledge, we need to pay special attention to existing information “pollution” and the need for information “purification”, for example by information selection and structuring. Information ecology can be helpful in that task and theoretical reflection on these issues is thus required. Such a reflection is the foundation for the considerations of information and knowledge ecology.

This article concentrates on information and knowledge ecology, which is considered to be one of the most recent fields within information research. I intend to present the subject matter, the research context, Polish and foreign researchers, current achievements, relationships in respect of other research fields, and then postulate on future directions of development. Those topics are discussed in the context of general trends in scientific development, as well as in the area of information and knowledge organization.

The goal of my article is to highlight the main problem areas of information and knowledge ecology as a field of research, which is presently considered to be one of the most promising directions in the development of information science research. This area shows clear signs of becoming an independent field of research, as indicated by the fact that some researchers have assigned it the status of a separate scientific discipline. This research project relies on the critical analysis of literature content following the comparative method.

**Information Ecology: from its etymological meaning to a metaphor**

We live not only in the era of information overload, but also in the time of real information pressure, even information violence, of which Piotr Chrząstowski wrote: “[…] information is merciless. It fills each gap it can penetrate, using every moment of our carelessness to encroach and occupy space wherever it can find some empty space.” (Chrząstowski 1997). However, the right to information should rather yield to the right of protection against information flood and intrusion. People find it easier to acquire unknown information unconsciously. Digging for information in Google, they usually have no idea how potentially useful information can be. They do not endeavour to acquire more information for that purpose. In fact, there is information we know that we know, and information we know that we do not have an idea of.

In information science, the term “knowledge organization” (often called information organization) refers to an organization of messages contained in documents. These messages represent knowledge or information. Such an understanding of knowledge organization is based on the assumption that each message within a field is a knowledge message. Knowledge contained in documents requires a communication process,
appropriate infrastructures and organization, and quality evaluation.

Information ecology is a metaphor treating information space as an ecosystem or infosystem. These terms express the relationship between conceptions of the natural environment, the dynamics of development, and the features of natural and digital information spaces. When describing and analysing the information environment, including information systems, information ecology uses the language of ecology as a science referring to the natural environment (or nature itself). Information ecology is a research conception postulating the study of information processes (information collection, organization, studying and making information available), in a manner that is analogous to the study of ecological processes conducted by biologists. What is important in this area is research into the factors influencing the condition of man’s information environment (anthropoinfosphere) and the influence of the information environment on the human psyche.

Excess of information, reflected in information overload, to give an example, is one of the reasons for the so-called information crisis, resulting, for instance, from the unbalanced development of various human spheres: intellectual, volitional, affective, and so on. Information ecology claims that the reasons for the ecological crisis lie in an unsustainable development of the information environment, reflected in the disturbance of proper proportions (or balance) between the factors that stimulate and control development.

The control over information and the information environment (and its status) is perceived as a natural phenomenon, in opposition to information domination over man, which occurs precisely during an ecological information crisis. That is why information ecology proposes a vision of man and society operating in harmony with the information environment (or ecosystem), adjusted to one’s natural capabilities. Consequently, there is a possibility reducing the split between a natural state of anthropoinfosphere and artificial and unnatural information production in the human environment, including the “crazy” generation and publication of a large quantity of “information” trash, causing information environment pollution (Sutter, 1998) and even “information diseases” (Bawden 2009). Recent infoecological approaches increasingly tend to reject the opposition between the processed (or artificial) environment, created by the Internet, and the natural environment, instead seeing the relationship of the natural and virtual environments as one of “complementation”, concerned with preserving the balance between the two.

Bonnie Nardi and Vicky O’Day define information ecology as “a system composed of people, actions, values, and technologies concentrated in a single local environment.” (Nardi, O’Day 1999). Information ecology emphasises and studies the influence of environmental information factors on man and vice versa. Consequently, information ecology concerns the relationships between man and his information environment. An infoecological approach thus relates to the search, within man’s
information environment, for those elements and interrelationships that indicate how
information affects man and vice versa, including in particular, the protection of man
against harmful influences of information and the protection of information itself
against the destructive actions of man. The information environment is, more precisely,
an ecologically determined complex of information and communication relationships
(Babik 2014, p. 59).

The context of the information ecology research field is determined by the following
issues:

– Identifying qualitative and quantitative information criteria;
– Studying the relationships between information and human health;
– Studying information values;
– Studying the processes of information transmission and reception;
– Studying the quality of information services;
– Determining liability for information and the social consequences of information;
– Studying the development and evolution of the information environment;
– Managing information at work, at home, and in society (Eryomin 1998).

Within this definition of the research field, what constitutes the object of information
is the ecology information environment, together with human information needs,
information barriers, information behaviour, information culture, information ethics,
information consumption, information diseases, information prevention measures,
information hygiene, information safety, information policy (Babik 2014, p. 110),
information ecology of particular fields, and ecological information management,
including its “ecological” organization.

Within information ecology, we can clearly distinguish two research trends:
diagnostic (pessimistic) and prognostic (optimistic). The former is associated with
diagnosing the present condition of the information environment and its ecosystem (or
clarifying the causes of the present condition and identifying future consequences). The
method of diagnostic surveying is primarily applied here with the intention of
identifying any existing threats. The latter trend is associated with forecasting human
development and human and social information environments. The authors
representing the latter trend primarily apply the methodologies of projection,
forecasting, and shaping the human information environment.

A key task of information ecology is the protection of the human information
environment, which is threatened mainly by modern information and communication
technologies. The means of fulfilling that task include the following:

– Concern for information quality (in the interest of man since non-organized
information and low-quality information affects people negatively), with respect to the
area that constitutes actual or virtual information space.
– Prevention of generating information that does not fulfil organization and quality
standards, which can become a specific “poison” polluting the human and business
information environments.

- Removal of information that has permanently and completely lost its usability, as well as disposal of information, and recovery of information usability if possible.
- Concern for valuable, safe, and well-managed access to organized information and knowledge, with the removal of information barriers.
- Shaping the human information environment.
- Initiation of the improvement and effectiveness of human operation in the information environment, including research on the processes of information and knowledge perception and reception.
- Bringing the status of sustainability and harmony to the anthropoinfosphere in order to ensure sustainable development of man in the information and knowledge society.

Practical actions within information ecology are concentrated mainly around activities intended to increase people’s information awareness in information processes, protection of people against becoming objects of information (or objects of manipulation), development of people’s information competences allowing them to manage information reasonably, control over excess of information, becoming independent of undesired information influences, especially of manipulation of people’s attitudes and behaviour; education towards responsibility for creation or generation, organization, processing, distribution, and use of information, and balancing human development in the world of technology and information.

Information ecology pays special attention to the information flood resulting from the human brain’s limited volume, the information environment of the information generator and receiver, the need to voice potential threats in the information sphere, as well as formulation of recommendations concerning information and knowledge organization, and the sustainability of information processes in education, the need for skilful use of information in building both individual and collective knowledge and creation of a personal information environment.

Information ecology thus proposes a vision of man and society’s functioning in harmony with the human natural information environment, that which is adjusted (or adapted) to natural (and sustainable) human perceptive capabilities (e.g. those of both students and teachers). The discipline pays attention to the alarming split between artificial and unnatural production of information in the human environment, including frenzied generation of the maximum possible “information trash,” causing pollution of the information environment and “information diseases” on the one hand and the natural condition of the anthropoinfosphere on the other hand. The discipline also proposes clear principles of generation, organization, distribution and use of information, disseminating the principle of information sustainability in compliance with “natural information transmission”.
Information ecology as a challenge for research on information and Knowledge Organisation

As was mentioned above, information ecology appeared only recently, in the 1970’s, although similar research had been conducted in the USA and Western Europe since the early 1950’s. Many publications considerably anticipating the emergence of the new field could be associated with the issues that currently belong to the area of information ecology (Xiwei et al. 2017). The field is treated as a separate discipline of knowledge studying the laws of information operation within so-called information ecosystems, as well as interrelated social, cultural, and political subsystems, including in particular the creation, organization, flow, and use of information by people in organisations and societies. That field of research deals with the mutual information relationships among people, events, technologies, and their information environments. Essentially, attention is paid to people and information and information processes.

Particular studies concerning such issues include the excess and shortage of information, information and knowledge organization, information barriers, information stress, information noise and smog, shaping information competences, and information niches. The infosphere allows the human information environment to be properly shaped, a process in which an essential role is played by information and knowledge organization. The new research field is interdisciplinary in nature because it draws its inspirations from a number of fields, such as information and knowledge management, human-computer interaction (HCI), law, and information science.

The specific research projects in our area of interest assume special significance in the context of the creation of an electronic information environment. Consequently, the information environment has reached such an excessive level of complexity in on-going interactions that it cannot make any predictions. In this context, the application of the concept of information ecology consists in integrating typical issues within information ecology in a dynamic digital information environment, requiring a specific organisation to provide necessary information and knowledge. Such an environment becomes more and more complex and “polluted,” or non-ecological, yet, at the same time, it cannot be ignored. There is a high intensity and scale of both desired and undesired phenomena in this environment. The Internet has made certain previously marginal phenomena massive and that is why they have become (or may turn into) social pathologies. A number of social barriers have been removed regarding means of accessing information, leading to both positive and negative consequences.

The following are mentioned as the pioneers of information ecology: Alvin Toffler (“Future Shock” (English ed., 1970)), Juliusz Lech Kulikowski (“Information and the World We Live in” (Warszawa: Wiedza Powszechna 1978)), and Neil Postman (“Technopoly: The Surrender of Culture to Technology” (English ed., 1992)). Information ecology as such is associated with the ideas and research of two Americans: Thomas Davenport and Laurence Prusak (Davenport, Prusak 1997), as well
as the Russian Alexey Eryomin (Eryomin 1998), who wrote about information ecology in the 1970’s in the USA and the former Soviet Union, respectively. Subsequently, that field of study was also developed by Rafael Capurro (Hochschule der Medien Stuttgart, Germany) (Capurro 1990) and Jela Steinerová from the Department of Library and Information Science, Comenius University in Bratislava, Slovakia (Steinerová 2012). There is also a small group of researchers in Poland involved in information ecology (among others Wiesław Babik, Ewa Głowacka, Wioletta Jachym, Katarzyna Materska and Beata Taraszkiewicz). Some of them adhere consistently to the phrase “information ecology” while others avoid that phrase for various reasons (among them Tomasz Goban-Klas, Marian Golk, Stanisław Juszczyk, Maria Ledzińska, Janusz Morbitzer and Ryszard Tadeusiewicz) (Babik 2014, p. 125).

Balancing the development of man’s information environment (anthropoinfosphere) is the main challenge of information ecology. It is an attempt at return to a natural balance between information excess and shortage. Ecological activities are intended to prevent various types of information anomalies in the information environment, and, as a result, they become a catalyst of human development. Information ecology puts stress on information quality and credibility, economy of messaging, need for metainformation, concern for the “purity of information,” and “clean messaging”, that is, proper information organization, to attain the goal of information flow without noise interference, or so-called “information stuffing”.

The idea of the sustainable development of the society of information and knowledge is an attempt to react to a complex of threats resulting from the uncontrolled (careless, irresponsible, and impetuous) generation of information and knowledge, with its negative influence on man. Sustainable development is the kind of development that does not interfere with natural eco(info)systems and networks that constitute the anthropinfosphere in which man operates. In the opinion of information ecologists, an antidote to the current information maladies can be found by using the same voice of reason and related moderation which are essential to the idea of sustainable development in many fields of human life and activity. There can be no overdose of common sense. For this reason, researchers stress the need to protect man against the excess of information and the importance of being concerned about information, including proper information and knowledge organization.

Information ecology accommodates the current problems of the broadly understood communication process and offers a solution conducive to process optimisation, in accordance with the needs and possibilities of information users. The researchers consider it indispensable to apply the principles of proper organization, prevention, hygiene, and a specific type of information diet, as well as the need to predict the consequences of one’s decisions in the area of influencing the information homeostasis of one’s own being and that of another person. What is of key importance in actions towards attaining information ecology is a change of mentality, shaping attitudes, and
building related social awareness.

Information ecology brings a new approach to information. The discipline is anthropocentric. Consequently, when exploring infoecological problems, one learns something more about oneself. An infoecological approach is, in fact, an inter-, multi- and transdisciplinary approach. Information ecology also assists man, understood as an information and cognitive system (Hetmański, 2008), when one defines man’s place in the real world, as well as in the world of electronic media (or virtual reality), according to the categories of opportunities and threats.

Therefore, it is necessary to follow the paths of neuroscience (Duch 2014). Such an objective should be accompanied by appropriate education supporting the system of humanistic values, intended to educate people to adopt a mature attitude in respect of information. It will thus be necessary to switch from information literacy to information culture (Batorowska 2013), or, in fact, to information ecoculture, supporting man, operating within the society of information and knowledge, by proper organizing, dosing and receiving information.

This area of information ecology research and the interests presented in general terms above also includes information and knowledge organization. Infoecological reflection is required in the organization of information and knowledge, as well as being a feature of information management, which, indeed, seems rather obvious. The most noticeable practical problems associated with information are the lack of scientifically justified standards concerning the organization and quality of information organization, and the lack of a clear and precise policy towards negligible and obsolete information, which essentially influences the processes of the organisation of information and knowledge.

The organization of information and knowledge constitutes a wide area of issues concerning ordering of information and knowledge to facilitate data search. Information explosion and, consequently, permanent development of information resources cause many difficulties in reaching information and knowledge. Information and knowledge are essential elements of man’s diverse and dynamically developing information environment. It is necessary to eliminate, or at least limit the existing barriers in the use of information and knowledge. Therefore, theoretical reflection on these issues is required. Such reflection is the foundation for considering organization of information and knowledge, facilitated by the existence of many different systems of organizing information and knowledge.

Conclusion

The continuous increase in the quantity of poor quality, fragmentary, and unreliable information means that it is imperative to implement reasonable processes of evaluation, selection, and proper organization of arriving streams of information (Babik 2016).
The use of information and knowledge ecology in knowledge organization for control and coordination of information and knowledge-generating processes may lead to the integration and creation of new added value. Ecological functions extended to the sphere of information processes also entail purely organizational aspects belonging to information and knowledge organization. These can lead, in particular, to the optimization of information resource structures, streams and flows. Integration of these processes may ensure the desired quality of information service for users, and such organization of information and knowledge as to allow them to obtain optimum levels of information use and access to new knowledge as added value. Information and knowledge ecology is beneficial for the optimization of the distribution of information which serves knowledge structure building. The use of information and knowledge ecology can enrich knowledge organization theory and systems.

This paper indicates that information and knowledge ecology allows for a wider and deeper view of knowledge organization and identification of the elements which have been previously omitted, or recognized to be insignificant. It also opens a new research perspective on knowledge organization and, consequently, a possibility of further development in the consideration of knowledge organization.

References


Elliott Hauser

Information from jussive processes

Abstract
This paper aims to augment a domain-independent conception of information processes developed by Losee (2011) such that it is philosophically pluralistic as well. If successful, I will provide a conception of information that can more easily be utilized across fields without requiring specific ontological or epistemological commitments unique to realism and empiricism. The information concepts of epidata and episemantics are introduced and operationalized in this context to bridge information processing and knowledge. These new concepts are related to Guattari’s hybrid semantics. Finally, I detail the relationship between these concepts and disciplinary and professional values of pluralism as described by prior thinkers. This work is intended to support scholars and professionals who engage with academic disciplines which view themselves as incommensurable with others.

Introduction: Losee’s information from processes
Losee (2012) has argued for a view of information in terms of processes, where a process’s output is informative about its input and/or the process itself. This is a promising, powerful, and succinct framework, but hinges on two key theoretical definitions: those of processes and of aboutness. Losee stresses that his theory is intended to be domain-independent, yet he views both aboutness and processes in ultimately physical terms (Losee 2012, 27). This is operationalized in his conception of hierarchical processes, which always have physical processes as the lowest layer. As an example, Losee’s representation of the Shannon-Weaver model of communication is shown in Figure 1. Losee’s theory sees hard sciences such as physics or chemistry having much more fine-grained colligations, since the processes they study are ‘closer’ to physical bedrock, while the social sciences such as economics or sociology must utilize a relatively ‘tall’ hierarchy of processes, making their disciplinary information more diffuse in physical terms. Thus, in Losee’s framework, the more precise hard sciences would seem to be more amenable to the methods of information science than the social sciences. If true, this would be a dismaying result for practitioners serving patrons in other fields.

Due to the priority given to physical processes, Losee’s information from processes framework as originally stated entails an epistemology of logical empiricism and an ontology scientific realism; it can most fruitfully be applied to domains that endorse these philosophical perspectives. Though indeed domain independent, this framework is ultimately disciplinarily reductive, and needlessly so. In this paper I’ll explore alternate definitions of process and aboutness to determine whether an alternate conception of them more widely compatible with other philosophical presuppositions is possible.
Figure 1: Losee’s representation of the Shannon-Weaver model of communication as hierarchical processes. The bottom of hierarchical processes is always physical.

Adapted from Losee 2012, p. 27.

**Jussive processes: encoding and interpretation**

Information processes are best described by what Bowker (2005), following Derrida (1998), has termed their jussive qualities: the things they leave out (Couldry and Hepp 2017, 129 et seq.). The jussive mood is a linguistic concept indicating a command, and in this context indicates that the active and contingent process of preservation exerts control over what ends up in the *formal archive*.

Bowker claims there are two distinct types of jussive processes at work in the formal archive: ingestion and retrieval (Bowker 2010). That is, the formal archive conditions remembrance by virtue of what is included within it and what is subsequently taken from it. In the context of information systems, I’d like to slightly reinterpret these two processes as *encoding* and *interpretation*. This allows a new definition of Losee’s information process as a jussive encoding, restrictive in terms of what and how it produces. The result of jussive encodings I label epidata, which isn’t ‘about’ anything; it’s merely, and materially, ‘subsequent-to’ the jussive process which create it. The the interpretation of a jussive process’s output I term episemantics, or the interpretive production of meaning from epidata. Episemantic interpretation is shown to be jussive in Bowker’s second sense, and it is only after such interpretation that aboutness can be said to obtain. This definition related to the Shannon-Weaver model of communication in Figure 2, further discussed below.

My repeated use of the prefix *epi-* in the terms advanced in this paper is deliberate. Defined by the Oxford English Dictionary as “upon, above, in addition”, it possesses the additive, active, constructive quality I wish to invoke. Jussive processes restrict, command, and create. The results of these processes are *derived from* or *in response to* their inputs but do not necessarily share any particular attributes or similarities with them. Further comments on each new term will follow.
Encoding and epidata

Encoding, viewed as a jussive process, restricts the content of a source as it enters a system by restating it within the system’s structure, schema, or terminology. For instance, a typed message on a computer is a jussive encoding of the act of typing on a keyboard into a binary character encoding. The keyboard and other parameters of input constrain the nature of the encoding. The material representation of binary, usually electrical and/or material potentials, are the material manifestation of this constrain. As Losee anticipated, this conception may be fruitfully employed to characterize chains of processes. For instance, the act of typing on a keyboard may be viewed as the jussive output of the process of writing on the writer’s thoughts or intent. Subsequent chains of encoding will materially process this text into electromagnetic charges on a hard drive, signals across the Internet, charges within RAM and cache, and variously illuminated pixels on a display. When a reader encounters the text, all they have access to is what was preserved in this chain of jussive processes. Though the text may result from a writer’s thoughts or intent, manifested in their input activities, the thoughts or intent are left behind when they became encoded. Interpretation, discussed below, is a process of reconstructing or attributing an epistemic dimension to this epidata.

Continuing the example of typing a message, it is clear that Losee’s formulation of information from processes assumes that the process of interpretation is unproblematic. In Losee’s framework, a process’s output (i.e. a typed message) is informative about the process (typing) and its inputs (the typing motion). This interpretation works well when interpretation is unproblematic: useful information about what keys a user pressed can be gleaned from computerized text. But in the case of how such keypressing related to the writer’s thoughts or intent, the interpretation is wanting for those who would seek to problematize the relationship between text and authorial intent. Keypresses in a word processing program can be said to be transparently jussive; that is, we value them precisely because of their unambiguous connection to their output. Interpretation is not problematic when we focus exclusively on the materiality of input. But the encoding of thought into words is a significantly more complex problem. Ambiguity is pervasive in this relationship, and Losee’s dictum (“a process’s output is informative about the process and its inputs”) seems to break down, precisely at the point where he uses the term “about”. It seems impossible for ‘aboutness’ to apply at the level of encoding or epidata. This is precisely what I’d like to claim: that additional theoretical work must be done to make space for this problematic, which is central for many academic disciplines.

Interpretation and episemantics

I’d like to propose that the solution to this problem is reserving the concept of aboutness as the result of a process of interpretation. Interpretation, you’ll recall, is my term for the second of Bowker’s two jussive processes, where information is taken from
an archive. It is the ascription of meaning to some signal, a process that I’d like to view as constructive rather than deductive. Continuing the example of encoded text, a reader ascribes meaning to the words on the page. In this act, the subsequent-to relationships of encoding are passively taken to be in a relationship with their inputs, as described above, but the additional construction of meaning by the reader represents a process different in kind from encoding. Only at the stage of interpretation can aboutness be coherently examined.

The term I’d like to introduce for this relationship is episemantics. The epi- prefix is intended to serve as a reminder that the semantical relationships here do not inhere in the encoded epidata but are rather placed upon, around, and in addition to such encodings.

Figure 2: An modification of Losee’s interpretation of the Shannon-Weaver model showing epidata, episemantics, aboutness, and subsequent-to relations. Aboutness relations are only possible after episemantic interpretation has occurred.

Adapted from Losee 2012, p. 27.

**Episemantics and aboutness**

With the distinctions between encoding/interpretation and epidata/episemantics in place, I will argue that Losee’s conception of aboutness’s role arises from a category error: while processes’ output is related to both their input and the processes themselves as he claims, that relationship should not be described as aboutness until episemantic interpretation occurs. Following logical empiricism, Losee assumes that episemantic interpretation is (or: can be; should be; for science, must be) a transparent process, enabling processes’ outputs to be about their inputs. I contend that aboutness only obtains in the relationship between the interpretation process and the jussive encoding process. The relationship between the epidata inputs and outputs of information processing should be described, weakly, as something like ‘subsequent-to’. The aboutness relationship consists of and is created by the episemantics of interpretation.

The major difference between Figures 1 and 2 can be seen as a re-centering of the process of meaning making from the Source to the Destination. In Losee’s model, the
line between Source and Destination indicates that meaning is transmitted from Source to Destination. We can label this conception of meaning making as *given-as* by the Source. In contrast, Figure 2 emphasizes that aboutness is a process of *taking-as* by the Destination. Losee’s conception of meaning transmission is more suited to the *subsequent-to* relationship that holds between epidata and the information processes that generate it. Losee’s conception of communication can thus be seen as the identification of what a message is *given-as* with what it is *taken-as*. My revision of his conception is intended to make space for the problematization of these processes employed by many disciplines.

**Aboutness and signifying and asignifying semiotics**

Episemantics and epidata in this sense can be shown to be similar to Guattari’s distinction between signifying and asignifying semiotics (Genosko 2016, chapter 1). While a full exposition of Guattari’s semiotics is outside the scope of this paper, I hope to show that the moves I am making have precedent. Further, this connection to a philosophy vastly different from and even (in the eyes of some) hostile to scientific realism is further evidence for the pluralism of my abstraction of Losee’s work.

Guattari’s semiotics is based off of an interpretation of the semiotics of Hjemslev (Guattari 1984, 73 et seq.). A crucial distinction for Guattari is Hjemslev’s separation of (roughly, physical) matter and the *formation of semiotic substance*. For Guattari, the relationship between semiotic substances and matter is an *asignifying semiotics* that Hjemslev overlooked. This term indicates the material component or effects of semiosis, a materiality that can have conditioning or constraining effects on subsequent semiotics of signification. The electrical, magnetic, and luminous components of a computer are all examples of asignifying semiotics. An emphasis on materiality has become pervasive in media theory, and much of it has a basis in Guattari’s work, both individually and in collaboration with Gilles Deleuze. The outcome of this focus has been to emphasize the impact of the materialities of media upon what can be and is expressed within them.

_Semiologies of signification_ are the more familiar acts of meaning-making, the things that are expressed within systems. But, for Guattari, they are also a tool of State power.

Writing machines are essentially linked to the setting-up of State power machines. The moment they are there, all other poly-centred semiotic substances become dependent upon a single specific stratum of the signifier. The totalitarian nature of that dependence is such that, by a tremendous retroactive effort, it seems to make all semiotics originate from the signifier (Guattari 1984, 75).

For Guattari, semiologies of significations’ power consists in their ability to crowd out all other modes of semiosis. Chief amongst these is the primitive mode of _symbolic semiologies_, which are employed in ritual and childhood play (Guattari 1984, 74).

In Guattari’s terms, everything from the encoding through transmission in Figure 1 is asignifying. This parallels my contention that aboutness does not yet apply. Only
after episemantic interpretation, which parallels Guattari’s semiology of signification, can aboutness obtain. This stands to reason, since without signification and its resultant semantics, there isn’t yet anything for aboutness to apply to. In terms of Guattari’s reading of Hjelmslev, aboutness must wait for the formation of semiotic substance.

Just as we were able to abstract away from Losee’s logical empiricism, we need not follow Guattari in seeing signification as an inherently political process. But, crucially, the framework of encoding and interpretation can accommodate Guattari’s formulation of semiotics. Asignifying semiotics can be seen as an analog of the jussive encoding of epidata. And semiologies of signification can be seen as the analog of the jussive interpretation of episemantics. Whereas, in Losee’s rendering, interpretation is non-problematic and, for science, should be transparent, for Guattari this becomes a locus of political power. The jussivity of epidata is the site whereby asignifying semiotics conditions subsequent signifying interpretation, and the episemantics of interpretation is the site of political control via signification.

Finally, I’d like to argue that the connection of these ideas with a thinker like Guattari, who is extensively discussed in critical and media theory, is evidence for the claim that my modification of Losee’s framework is indeed philosophically pluralistic. Guattari has been sarcastically described by physicists as “the most brilliant mélange of scientific, pseudo-scientific, and philosophical jargon that we have ever encountered; only a genius could have written it” (Sokal and Bricmont 1998, 166). Simultaneous accommodation of scientific realism and Guattari’s semiotics is, I offer, pluralism in the extreme. It also indicates the stakes of my project. If information science is to serve diverse intellectual constituencies, some of which are fundamentally dismissive of others, pluralism is the only strategy that can reconcile these diverse needs without operationalizing a normativity against avowedly incommensurable disciplines.

An Ontologically pluralistic, domain-independent framework for information

Together, these modifications seek to transform Losee’s theory from one that entails and requires assent to scientific realism to one that is more philosophically pluralistic, while still preserving domain independence (i.e. psychological processes producing psychological information, biological processes producing biological information, and so on). Further, it provides a framework to characterize different domains and systems by describing their jussive processes and the interpretive practices they enable. Losee’s original construction can be seen as a particular instance of my more abstract formulation wherein jussive processes are equated to some colligation of physical phenomena and interpretation is assumed to be unproblematic. The alternate constructions enabled by this abstraction open it to a much broader application by precisely characterizing these elements in distinct ways.

Systems, viewed as the assemblages of various jussive elements, can be described in terms of the epidata they produce, subsequent to various inputs. Users of systems
can then be seen to place episemantic interpretations upon systems’ epidata, producing meaning via these aboutness relationships, which are often heavily influenced by disciplinary and cultural particularities. This conception suggests that a core benefit of contextual techniques like domain analysis can be described as a focus on enabling users’ preferred episemantic interpretations via the specific jussive processes employed in system design. It further suggests that the cognitive authority of information organization professionals (Wilson 1983; Mai 2010) may be directly related to their ability to accomplish this enabling. This restatement of pluralistic values has potential applications in contextual design methodology, the creation and use of controlled vocabularies, interdisciplinary retrieval systems, and interoperability of systems.

Conclusion: philosophical pluralism in Information Organization

Why am I interested in doing this work? Amidst long-standing calls for pluralism (Wilson 1983; Mai 2010), and the development of contextual design methodologies such as domain analysis (Hjørland 2002), the field needs frameworks with both domain independence, which Losee provides, and epistemological independence, which I attempt to provide here. Together, these independences will deepen the field’s ability to serve a wide range of domains holding a wide range of philosophical presuppositions.

I don’t seek to claim that Losee himself ignores pluralism as a goal. In fact, Losee makes important overtures to disparate philosophies, such as when he shows how both coherence and correspondence theories of truth can be represented as processes (Losee 2012, 162–165). While these overtures show that Losee’s project is motivated by a species of pluralism, I’ve argued that fundamental modifications are required for true success here.

I’ve shown how Losee and other scientific realists can continue to use and advocate for their preferred view that the jussive elements of processes should be described physically, that episemantics is an automatic, or at least unproblematic, process, and that subsequent-to relationships are sufficient for attributing aboutness. These claims will be unacceptable to many other users and domains, however, necessitating the approach in this paper. Ultimately, my goal is to extol and commend the unifying power of Losee’s approach while advocating for an abstracted conception of it suitable for the inherently pluralistic nature of information science.
References
Joseph T. Tennis

Intellectual history, history of ideas, and subject ontogeny

Abstract
Outlines the general methodological stances in subject ontogeny research through examination of work in intellectual history, history of ideas, and in the philosophical commitments of realism, conceptualism, and nominalism. From this we can construct a rich vocabulary that allows us to differentiate kinds of subject ontogeny work through these descriptions of methodologies.

Introduction
In the context of subject ontogeny research, that is, the study of how subjects change or do not change through revision of indexing languages, there a number of methodological questions. The nature of semantics, ideas, and subjects is a long and rich discussion in the context of knowledge organization (Ranganathan 1967; Wilson 1968; Foskett 1969; Hjørland 1992; Mai 2001; Olson 2002; Zeng, Zumer, and Salaba 2010; Dutta 2015). This discussion is kept current by Birger Hjørland’s reference work on the matter (Hjørland 2017). However, the discussion predates the founding of the discipline of library and information science. Ranganathan cites the early Vedic work (fl. 1500-600 BCE) on subjects (Ranganathan, 1957). Plato (c. 380 BCE) lists those subjects that are required for a good education of future leaders and contemplates their nature (Emlyn-Jones and Preddy 2013). Following Plato’s dialogues the Western encyclopedists, and ultimately bibliographers, write subjects into their work, sometimes reflecting on them, but often not (c.f. Diderot and d’Almbert 1751-1772).

Our contemporary preoccupation with subjects and how they remain the same or change is linked directly to our philosophical assumptions. In this context we often talk about two extremes: realism and nominalism (Boswell 1986). Realism considers subjects (categories, ideas, semantics) to be “the footprints of reality… they exist because humans perceive a real order of the universe and name it” (Boswell 1982 p.91). The nominalists subscribe to “the belief that categories are only the names… of things agreed upon by humans, and that the ‘order’ people see is their creation rather than their perception” (Boswell 1982 p. 91). The middle way, conceptualism, places universals and abstract objects in our minds, and leaves the relationship between these concepts in the mind and the world outside our minds as a matter of debate or camp (McLear 2014).

There are refinements and caveats to these two extremes, including a middle way in conceptualism (cf. McDowell 1994; McLear 2014). Further there are different and various ways realism or nominalism manifest (e.g., in the mind, external to the mind, in action, or in culture). Proponents of nominalism do not admit to the existence of
universals or abstract objects where “a universal is something that can be instantiated by different entities and an abstract object is something that is neither spatial nor temporal” (Rodriguez-Pereyra 2016 no page numbers given).

There are nuances to these camps. It is possible to hold a realist position and assume the same phenomenon is named differently in different context and in different points in time. And here is where the methodological question comes back to subject ontogeny work. Namely, does research done in one or the other camp use different methods to arrive at a description of how a subject changes over time in the life of an indexing language?

Because of our rich tradition within knowledge organization, and in the field of philosophy before, we can work within those contexts to find some answers, but to date, most of those discussions assume a kind of non-diachronicity to subjects, i.e., the reason we would change our indexing languages, to discuss subjects, would be to add new subjects, or to further refine the subject that was already present in the indexing language. The subject itself does not change in this case. It remains the same or is further specified. It is operationalized as a universal and abstract object as identified above.

However, we have a recent literature that looks at change (Fox 2016) or the lack thereof (Lee 2016), and builds on more than a decade of investigation into this topic (Tennis 2002). So if change is observed, what is changing? Is it the subject or the name or some admixture of the two? What ways can we make arguments for any of these three camps? This brings us to our discussion of the disciplines of intellectual history and history of ideas.

In a piece from 2012, Peter Gordon addressed the relationship between intellectual history and history of ideas. Of the former he said, “[b]roadly speaking, intellectual history is the study of intellectuals, ideas, and intellectual patterns over time. Of course, that is a terrifically large definition and it admits of a bewildering variety of approaches” (Gordon, 2012, no page numbers given). For this history of ideas, he says, it is a discipline “which looks at large-scale concepts as they appear and transform over the course of history. An historian of ideas will tend to organize the historical narrative around one major idea and then follow the development or metamorphosis of that idea as it manifests itself in different contexts and times, rather as a musicologist might trace a theme and all of its variations throughout the length of a symphony” (Gordon 2012, no page numbers given). He goes on, “[b]y insisting that the idea is recognizably the same thing despite all of its contextual variations, the history of ideas approach tends to encourage a kind of Platonist attitude about thoughts, as if they somehow preexisted their contexts and merely manifested themselves in various landscapes” (Gordon 2012, no page numbers given). In contrast, intellectual history “regards ideas as historically conditioned features of the world which are best understood within some larger context,
whether it be the context of social structure and institutional change, intellectual biography (individual or collective), or some larger context of cultural or linguistic dispositions (now often called ‘discourses’)” (Gordon 2012, no page numbers given).

In the history of ideas we see ideas as universals and abstract objects. In intellectual history we see particulars and spatiotemporal objects. As in history, in classification we rely on documentary evidence to construct a scheme for classification. This scheme can be seen as a document with rhetorical force (Feinberg 2011). Following Feinberg, one could argue that either explicitly or implicitly a scheme for classification is going to argue that subjects are universals and abstract objects or they are not. The work of subject ontogeny can either do justice to this authorial voice or not. What then, do the methods and results of subject ontogeny look like if they admit to different philosophical and by extension different historical views on subjects in long-lived schemes?

**Realism in subject ontogeny work: subjects as history of ideas**

As mentioned above, we have a rich history of discussing the nature of subjects in knowledge organization. Both materialist and rationalist conceptions of subject seem to share the idea that subjects are universals and abstract objects represented in schemes for classification. This seems counterintuitive, because the two are often presented as divergent epistemic stances in the literature (Hjørland 1992). However, if we take a view that the scheme for classification is a document, and we approach the reading of the document from the perspective of the history of ideas, we will need to be able to point to the same subject throughout time. That subject, as it is represented in the scheme is an *index* in the Peircean sense (Atkin 2013). That is, the class points to, or *indexes*, the subject. The subject can be postulated to be a universal or abstract object justified by an appeal to rationalism or the material nature of the subject.

In considering the adoption of this method of examining the ontogeny of a subject in a long-lived scheme for classification, we may or may not need to justify what we identify (index) as the subject. Our assessment of the identity of the lexical form in scheme may be enough. This has been the way Tennis (2012) has established his ontogeny of Eugenics in the Dewey Decimal Classification.

Like Gordon’s take on the history of ideas, the realist view of subject ontogeny work does not need to contextualize the subject beyond how the scheme characterizes it in relation to other subjects.

The caveat here might be in a materialist view of subject that, over time, is linked with various volumes in the literature. In this case we may see the subject, while constant, manifest in different texts. As noted by Tennis (2007), this could constitute a form of scheme change that may or may not reflect the ontogeny of the subject. While still realist in nature, if we consider the literature associated with subjects in schemes
for classification we move, ever so slightly, toward conceptualism or nominalism in subject ontogeny work.

Nominalism in subject ontogeny work: subjects as intellectual history

If nominalism is defined by the rejection of universals and abstract objects, then in order to observe change over time, classification theorists must contextualize the subject – not only in the scheme, but using other sources of evidence to follow their commitment to the fact that subjects, in the nominalist view, are particular and linked to other factors in space and time. In a radical appeal to nominalism, we might see research arguing that there is no single subject that changes through scheme revision. Following this each new scheme for classification, each new document, should be interrogated in its user and literary warrant context as well as within the structure of the scheme for classification. The researcher in this vein would need to bring in outside literature, and compare different slices of time when the subject was used to class literature.

This is not to say that a narrative about the subject could not be written to tell a story about how it moves from one version to another, but in each of these migrations, we would expect to see a recontextualization of the subject that might even require a redefinition. If we follow the intellectual history mode of inquiry, we would specifically call on the context of 1930s Indian civil engineering to interpret that subject in the earlier editions of the Colon Classification.

This mode of inquiry would precipitate a more robust description of what the researcher understands the subject to be in a given particular, spatiotemporal context. We would need a thick description – a categoriography of the subject and the class that represents it, in combination with literature and a particular kind of intellectual history that runs parallel to the time and space of the scheme for classification. This is especially true if we believe it to be a document with authorial voice and potential rhetorical force.

Conclusion

Subject ontogeny is an exciting and fruitful form of inquiry into many aspects of classification theory. Not only has it pointed to new design requirements for schemes for classification (and other indexing languages), but it also has laid bare some assumptions we have made about subjects and their stability. In this paper we have tried to point out at least two ways in which we can discuss change when we engage in a description of a subject’s ontogeny in a long-lived scheme for classification. We have more we can document, specifically on middle-way approaches like conceptualism. Further, while we may be able to continue with the methods we are currently aware of, more robust methods may need to be developed if we need to document context beyond the scheme.
References


Isidoro Gil-Leiva, Mariângela Spotti Lopes Fujita, Pedro Manuel Díaz Ortuño, Daniela Majorie dos Reis

Is the massive incorporation of e-books into university libraries devaluing the technical processes related to the assigning of subject headings and classification codes?

Abstract
In recent years, academic libraries have incorporated huge numbers of e-books into their catalogues. The increasing use of this new technology is having a significant impact on the technical processes involved in incorporating e-books into catalogues, as well as causing changes to library-user relationships and to the maintenance and management of collections. With this study, we intend to find out how e-books are affecting academic libraries in Portugal, Spain, the United Kingdom, the United States and Brazil, particularly with regard to the assignment of subject headings and classification codes. In order to gather this information, a web survey was prepared to obtain the opinions of the directors or technical process managers from nearly 500 academic libraries.

1. Introduction
University libraries are constantly incorporating numerous collections of e-books. A report by the Society of College, National and University Libraries (SCONUL 2015, p. 2) states that “e-books accounted for 24% of total book stock, on average, in 2013-14, compared to 11% in 2009-10”. In Spain, sales of e-books in 2016 experienced an increase of 41% compared to 2015 while in Latin America they rose by 110% (III Edition of the Spanish Markets Digital Evolution Report 2017, p. 3).

The following is an illustrative example of the incorporation of e-books in Spanish university libraries:

<table>
<thead>
<tr>
<th></th>
<th>2013</th>
<th>2014</th>
<th>2015</th>
<th>2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spain</td>
<td>29,858,894</td>
<td>30,029,980</td>
<td>31,008,371</td>
<td></td>
</tr>
<tr>
<td>Print books</td>
<td>28,718,064</td>
<td>(+4,36%)</td>
<td>(+3,15%)</td>
<td></td>
</tr>
<tr>
<td>E-books</td>
<td>8,819,673</td>
<td>10,767,685</td>
<td>11,955,167</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(+6,69%)</td>
<td>(+12,21%)</td>
<td>(+9,93%)</td>
<td></td>
</tr>
</tbody>
</table>

The emergence and expansion of e-books in libraries in the mid-2000s has led to talk of them being a disruptive technology or disruptive innovation (Frederick 2016, Buschow 2014, Wilson 2014 and Herrington 2013). The application of this new technology is causing major changes. Not only does the e-book raise questions about the survival of the printed book, but perhaps also librarians are faced with a possible paradigm shift in terms of their traditional way of working since different ways have
emerged to incorporate e-books into catalogues, involving alterations in technical processes, new forms of user-product relationships and modifications in the maintenance and management of collections.

Since the emergence and expansion of digital documents and their increased use for information browsing and searching in contrast to printed material, the future viability of indexing or controlled vocabularies has been questioned (see the special issue of Knowledge Organization, 2016, 43 (3) and in particular Dextre (2016)). Moreover, something similar has occurred with classification if we pay attention to what Hjørland (2012) indicates, namely that it is necessary to rethink classification to ensure its survival. Therefore, do e-books pose a new threat to the traditional librarians’ task of assigning subject headings and classification codes to e-books?

The incorporation of e-books into university libraries has meant the emergence of a significant production of literature related to printed book versus e-book, dealing with the issues of incorporating e-books into collections and how best to do so, in relation to their technical management, users' use and acceptance, as well as the problems, changes, challenges or even frustrations of librarians. We can thus offer an (albeit incomplete) outline of some examples of this literature.


We have not found publications focused on further analyzing the issue of assigning subject headings and classification codes in e-books. Hence, the main objective of this study is to know what is happening in academic libraries in Portugal, Spain, the United Kingdom, the United States and Brazil with the technical processes undertaken with respect to e-books regarding the assignment of subject headings and classification codes.

2. Materials and methods

To achieve the objectives set out in this research we will gather information and opinion from 500 academic libraries (Portugal, 20; Spain, 64; England, 160; United States, 120 and Brazil, 126) through their directors or heads of technical processes. The selected academic libraries from Spain, Brazil, England and Portugal are
practically all those that exist, while those from the United States can be considered a representative sample. The corresponding contact emails were readily available.

We have built a survey consisting of open-ended questions, closed questions and matrix questions, elaborated using the "Survey" platform of the University of Murcia to obtain information on possible changes in the processes of assigning subject headings and classification codes, the usefulness of metadata provided by publishers, the suitability of introducing automatic processes in assigning subject headings and classification, or as to whether we are facing a possible paradigm shift in relation to the technical processes involved with e-books, among other aspects.

The survey was conducted in English, Spanish and Portuguese, and an invitation was sent by email to the directors or technical process managers of almost 500 university libraries in Portugal, Spain, England, the United States and Brazil to be completed through a web browser. Eighty completed questionnaires were obtained after three weeks, which represented a 16% response.

3. Results

3.1. Assignment of subject headings and classification codes to e-books

Subject headings are assigned to purchased or subscribed e-books by 41% of respondents, 36% assign them to some e-books, and 23% do not assign subject headings. On the other hand, it seems that the process of assigning subject headings remains the same for e-books as for printed books, with 88% of respondents saying that there are no differences.

Classification codes are not assigned by 53% of respondents, 25% do assign them, and 22% assign them only to some e-books. On the other hand, 81% of respondents indicated that the processes for assigning classification codes are the same as for printed books, while 19% indicated that it is not the same process.

The assignment of subject headings is normally reserved for individual e-books, purchased and with permanent links, and for the recommended bibliographies of lecturers. E-books can provide sufficient and quality subject headings or require assignment of subject headings for different reasons: absence, need for translation, use of automatic assignment tools, or because more specific subjects are required, or a thesaurus is used. At the same time, classification code assignment is also reserved for individual e-books and purchased property, just like subject headings. In general, they are not used since the effort of assigning classification codes does not compensate for their limited use, and they are not useful for the discovery and retrieval of information by the user.
3.2. Characteristics of metadata related to indexing and classification provided by publishers/suppliers

Regarding metadata relative to the subject headings of e-books provided by publishers/suppliers, 39% of respondents considered them adequate; 37% were unclear and 24% did not agree that they were adequate. Concerning the issue of whether subject headings were sufficient and not worth modifying, 43% of respondents agreed; 31% said they did not agree; and 27% disagreed. The same question was asked in relation to classification codes: 40% think they are adequate; 29% do not consider them clear; and 31% disagree. Furthermore, 46% consider them to be sufficient and not worth modifying; 26% do not consider them clear; and 28% think that they are insufficient and do need modifications.

![Figure 1: Metadata related to subject headings of e-books is adequate / sufficient](image1)

![Figure 2: Metadata related to classification codes of e-books is adequate / sufficient](image2)

3.3. Processes

50.65% of librarians consider that the high number of e-books purchased / subscribed in the library makes an adequate assignment of subject headings unfeasible, while 14.29% are undecided and 35.06 do not agree with this affirmation.
In relation to classification codes, 56% of librarians consider that the high number of e-books purchased/subscribed in the library makes an adequate assignment of classification codes unfeasible, while 13.33% are uncertain and 30.67% disagree with this affirmation.

On the other hand, we also wanted to know if the same number of subject headings are assigned to e-books as to printed books: 70% of respondents said that they assign the same number of subject headings to e-books as for printed books, 21.25% did not assign the same number of subject headings and 8.75% said they did not know. We also wished to find out if there have been any changes in the application of technical processes to e-books in relation to indexing and classification, and if those changes have been introduced in the library's procedures manual: 54.67% of respondents answered yes, 42.67% said no, and 2.67% did not know.

3.4. Authorship of subject headings and classification

55.13% consider that headings should be provided by publishers/suppliers, 20.51% do not have a clear idea, while 24.36% consider that to be a task for librarians.

54.05% say that classification codes should be provided by publishers/providers of e-books, 21.62% are unclear, while 24.33% think that it is the job of librarians.

3.5. Influence of metadata quality on the selection of publishers/suppliers

We asked if in the selection of e-book packages for purchase or subscription, the quality of the metadata related to the subject headings and the classification codes provided by the publishers/suppliers is taken into account. 50.65% consider that the quality of the metadata provided is not considered in the selection, while 40.26% believe that it is.

3.6. Automation of the assignment of subject headings and classification codes

27.63% indicate that it is not appropriate to implement the automatic assignment of subject headings. It is not clear for 22.37% whether it is helpful to introduce automation while 50% clearly prefer to introduce automatic assignment, although there are doubts about the quality and adequacy of the automatic assignment of subject headings, since it could be incorrect or inadequate. Some respondents consider OCLC’s FAST project to be interesting, although currently it only appears to provide subjects in English.

33.78% indicate that it is not appropriate to introduce the automatic assignment of classification codes, 22.97% do not know if it is appropriate to do so, and 43.24% clearly prefer to proceed with automatic assignment.

3.7. Paradigm shift

There seems to be no consensus on this issue. Respondents consider the volume
of e-books to be increasing and that a reorganization of the library system is required. It would be important for publishers and librarians to work together to develop a model that addresses all user service needs. On the other hand, if librarians work in publishing houses, indexing and classification would be more effective. The main change that should occur is for large publishers to think of the librarian as an essential professional.

Furthermore, we may consider that the access to an increasing amount of information (including e-book databases) through the traditional forms offered by library catalogues and discovery services, based only on metadata rather than full text, no longer fully serve the user, who wants to obtain the information, and not the book. The user ignores the format of the resource: it is thus the resource itself that someone may be looking for and must be discoverable.

**Figure 3: Paradigm shift**

<table>
<thead>
<tr>
<th>Yes.</th>
<th>No.</th>
<th>I don't know.</th>
</tr>
</thead>
<tbody>
<tr>
<td>41.56%</td>
<td>40.26%</td>
<td>18.18%</td>
</tr>
</tbody>
</table>

**4. Conclusions**

In accordance with the results achieved, the following conclusions can be established:

1) Most librarians assign subject headings to e-books following the same processes as for printed books; fewer than half of librarians assign classification codes, also following the same process as for printed books.

2) In relation to the subject headings and classification codes provided by the publishers/suppliers, there does not seem to be a consensus among the respondents on whether they are adequate and sufficient.

3) Most librarians have pointed out that due to the high number of e-books purchased (or subscribed to) by libraries, a proper assignment of subject headings and classification codes becomes unfeasible.
4) Most respondents consider that the subject headings and classification codes of e-books should be provided by publishers/suppliers.

5) It appears that the quality of metadata relating to subject headings and classification codes is not taken into consideration when buying or subscribing to e-books from publishers (suppliers).

6) Half of the respondents consider that the current environment is suited to the introduction of processes for automatic procurement of subject headings and classification codes.

7) It is not possible to establish if we are facing a paradigm shift.

Finally, in terms of how users encounter subject headings and classification codes provided by publishers or suppliers and regarding the records in which the indexing and classification is performed by librarians, it seems that there has been a devaluation of the processes relating to subject headings and classification codes in university libraries.

References


Knowledge Organization in editorial policies for titles, abstracts and keywords in JCR-indexed journals: an exploratory study in the areas of Information and Communication Sciences

Abstract
In the formal writing of articles, the title, abstract and keywords have been the essential elements of knowledge representation. However, authors themselves are nowadays responsible for electronically making available the document representation elements (title, abstract and keywords) of their papers. For that reason, the editorial policies established by the journals, especially guidelines for authors regarding standardization, style, structure, extension and format for writing and submission of their articles are particularly relevant. The purpose of this exploratory study was to observe and analyze guidelines provided for authors on writing their papers’ title, abstract and keywords in a representative sample of Library & Information Science (LIS) and Communication Science (CS) journals indexed in Journal Citation Reports (JCR) taking into account the first 8 titles of each one of the four quartiles. The sample consisted of 32 CS journals, out of a total of 82 (39%), and 32 LIS journals, out of a total of 85 (37.6%). A spreadsheet designed for data collection and processing was used containing 36 items grouped into four categories: identification data, guidelines for the title, abstract, and keywords of the scientific article. The categories established for each knowledge area were then analyzed by comparison in order to verify the use of specific aspects of knowledge organization and representation in editorial policies. The results reveal a scarcity of scientific literature, a tendency to standardize general indications and criteria, divergent behavior and little instruction in LIS and CS about titles, abstracts and keywords. The conclusion is that these aspects are more related to protocols of editorial management proceedings than to the intrinsic quality of documentary products associated with scientific organization and representation.

1. Introduction
Representation practices in knowledge organization are important for information access and retrieval, contributing to the intellectual and social organization of knowledge (Hjørland 2003). The production of academic scientific knowledge that needs to be formally or informally recorded has always been linked to the publishing world, which establishes the forms of publication and distribution of knowledge.

Knowledge visibility came to be measured from the effects of scientific production and publications on citation indexes by calculating the impact factor of scientific journals, their main communication vehicle. With the advances of the Internet in information and communication technologies, the means of production and dissemination have changed irreversibly, thus enabling the academic world to dominate publishing and distribution, which has given rise to a movement known as open access.

However, due to the impact factor and publishers’ control over knowledge, open access to important scientific journals is not allowed and requires subscriptions. Nonetheless, open access and subscription-based journals are now also available on
Google and Google Scholar, whose meta search engines provide a fast and broad search for scientific literature through titles, abstracts and keywords.

In the formal writing of articles, the title, abstract and keywords have been the essential elements of knowledge representation since the appearance of the first scientific journals. Intellectual access is provided by organizing document representations in catalogs, bibliographies, indexes and databases (Hjørland 2003). Therefore, these representations are determinants of the visibility and measurement of the intellectual and social impact factor.

Scientific journals are, at the same time, an indispensable channel of expression and reception for the creators and consumers of information (Delgado López-Cózar, Ruiz Pérez and Jiménez-Contreras 2007). Until the emergence of the Internet, access to articles was controlled by online information retrieval systems that had very specific rules for the normalization of titles, abstracts and keywords. In the current digital environment, a growing number of scientific journals, especially those published by academic institutions, are available via open access to online repositories. This means that the conventional publishing world and its professionals have minimal presence and little control over the elements of documentary representation - titles, abstracts and keywords – having once contributed significantly to their elaboration by using specialized techniques and tools. Search engines, such as Google Scholar, specialize in delivering academic and scientific content by means of retrieval strategies based on full texts and natural language. However, nowadays authors themselves are responsible for providing the document representation elements (title, abstract and keywords) of their papers, and metadata of scientific journals are available electronically. In this context, the editorial policies established by journals, especially guidelines for authors as to standardization, style, structure, extension and format for writing and submitting their articles are particularly important.

International standardization bodies (ISO, UNESCO), stylebooks or style guides of publishing companies (Oxford University Press, SAGE), universities, scientific societies and professional associations (American Psychological Association) perform the standardization of scientific publications. Thus, they contribute to the process of systematizing and evaluating knowledge communication and dissemination.

2. Objectives and methodology

Considering the importance of these issues, the purpose of this exploratory study was to observe and analyze some editorial policies and guidelines provided for authors on writing their papers’ title, abstract and keywords. Journals indexed in the Journal Citation Reports (JCR) of Librarianship and Information Science (LIS - Library & Information Science) and Communication Sciences (CS-Communication Science) were used as representative samples.

LIS was selected because of its research interest and its production of knowledge
with regard to journals in scientific communication and, essentially concerning indexing themes and summaries for elaborating titles, abstracts and for determining keywords in scientific articles. The analysis of guidelines for elaborating titles and abstracts and for determining keywords provided by journals in Information Science could, in principle, be considered as a parameter for other areas. CS was selected as an area of comparison because it belongs to the broad field of human science and because it contains a similar number of journals. A quantitative approach addressed a representative sample selected from the JCR taking into account the first 8 titles of each one of the four quartiles. The sample consisted of 32 scientific journals from the LIS area out of a total of 85 correspondents (37.6%) and 32 from the CS area out of a total of 82 (39%). The representative samples show, in descending order, an impact factor interval of between 4.180 and 0.154 in CS and between 7.268 and 0.022 in LIS. The total of 64 selected journals were found by means of Internet searches which allowed for identification of website addresses and access to information necessary for the research.

In the exploratory study performed, the editorial policies about the elaboration of titles and abstracts and the determination of keywords of each of these journals were observed and analyzed by means of guidelines items for the authors. To this end, a combination of qualitative and quantitative techniques was used.

In order to perform the quantitative and qualitative analyses, a spreadsheet was designed for data collection and processing containing 36 items grouped into four categories: journal identification data, guidelines for the title, abstract and keywords of the scientific article and, at the end, a field for notes and observations.

Both common and specific data are present in this model: a) identification of the scientific journal (title, web address, ISSN, ranking position, impact factor, type of access and standardization manual and style adopted); b) guidelines for the title of the scientific article (writing, translation, language of translation, section of the journal, indications as to the number and types of words, type of information content, subtitle, abbreviations, acronyms, symbols and formulas and style); c) guidelines for abstracts (writing, translation, language of translation, section of the journal, indications as to the number of words, type of information content, abbreviations, acronyms, symbols and formulas, style and inclusion of bibliographical references); and d) guidelines for keywords (writing, translation, language of translation, section, indication of the number of words and separation punctuation, vocabulary control and type).

Subsequently, the categories established for each knowledge area (Librarianship and Information Science and Communication Science) were analyzed by comparison in order to verify the use of specific aspects of knowledge organization and representation in editorial policies.
3. Comparative results of LIS and CS journals

3.1. Direction on guidelines and style guide

It appears that only 22.6% (n.7) of journals in the LIS sample present guideline or style guide. Specifically, in the 1st quartile there is one journal, in the 2nd quartile there are 5 journals, and in the 3rd quartile there is one journal. In terms of publishers, Taylor & Francis includes two journals while the remainder are from different publishers (Wiley, Elsevier, Emerald, Sage and Medical Library Association). When guideline and style guide are identified, there is no uniformity in denominations, but most of them refer to “Author guidelines”, “Guide for authors. Author information pack”, “Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals”.

The style guide prevails in 84.3% of the CS sample (n.27). In six of these journals, the style guide is identified as “Author Guidelines/Instructions for author/Submission guidelines” besides referring to APA 6th. In this group, there are two journals published by Sage, three by Wiley and one by Taylor & Francis, illustrating the widespread acceptance of APA rules. In addition, there are eight Sage journals that turn to SAGE Manuscript submission guidelines, evidencing a tendency to standardization in journals of the same publisher.

3.2. Guidelines for writing title

In LIS, 56.7% (n.17) of the sample present indications for writing titles, with five journals in each one of the first three quartiles and two journals in the 4th quartile. These indications are included in the generic sections entitled “Guide for authors”, in a special section designated “Title”, in “Format guidelines”, in “Manuscript submission” or “Manuscript requirements”. In this group of 17 journals, there are 10 that include guidelines about the number of words. These recommendations refer to titles of quite a short length, with two journals accepting titles of up to 50 words, while the remaining journals point to more reduced titles of 45 characters, including spaces, to 40 characters, or titles ranging from 6-12 words to 16 words. In terms of the type of words included in the text, there is only one journal with a specific guideline advising against the use of expressions like "Investigation of ..."; "Study of ..."; More about ..."; "... revisited". Two journals only address the subtitle, indicating that the reference “research project” can be placed in this section or that it can be used to specify the content of the paper. There are also seven journals which advise that the title must not contain abbreviations, formulas or references, in order to make it shorter but also focused on essential information. Taylor & Francis recommend authors to think about words which are relevant for their field of study. To increase visibility in information retrieval systems, it is suggested that authors use a key sentence, emphasizing the need for a concise, precise and informative title.
In CS, 71.9% (n. 23) of journals include guidelines for titles in the Author guidelines, Submission guidelines or Instructions for authors sections. In five cases, the section is entitled “Help readers to find your article”, emphasizing that there is concern in creating titles that work as elements that enhance the visibility of articles. However, there are few concrete guidelines for writing titles, with only two journals in which a limit of characters for titles is mentioned, in some cases indicating a short title of 50 characters and in other cases of 40 characters. There are also four journals that suggest authors should think about the research terms that readers can use to look for information. In this regard, specific guidelines are provided as with the use of key sentences which are easy to read, considering that users search for expressions or phrases and not only isolated words. It is also suggested that the title should be understandable for a reader outside the field in question and that, whenever possible, abbreviations, formulas and numbers should be avoided. In addition, authors are advised to avoid expressions such as "Investigation of ..."; "Study of ..."; More about …"; "... revisited".

3.3. Guideline for writing abstracts

In LIS, 76.7% (n.23) of the sample present guidelines for writing abstracts in the same guideline sections as those for titles. In this group, there are 18 journals that specify the number of words to be taken into account in the abstract which varies from 50 to 250 words. There is no concern in adapting the type of abstract to the article because only one journal points out that there are significant differences between original research documents and review articles. For original documents, the description of the method and proceedings for reviews provide a different approach: first, there must be an indication of the primary aim of the review, the reasoning behind the choice, the main outcomes and the outcomes from the review as well as the conclusions, including implications for new research, applications or practice. Furthermore, there is no evidence that structured abstracts are common in LIS journals because only four indicate the need to create such an abstract: one of the journals requires the indication of objectives, outcomes and conclusions; two journals consider the possibility of 4 to 7 sections, some mandatory and some optional: Purpose (mandatory); Design/methodology/approach (mandatory); Findings mandatory); Research limitations/implications (if applicable); Practical implications (if applicable); Social implications (if applicable); Originality/value (mandatory); while, the other journals require "Background", "Objective", "Methods", "Outcome", "Discussion" and "Conclusion". With respect to the use of abbreviations, acronyms, symbols and formulas, seven journals advise against and six journals do not permit the use of bibliographic reference. In terms of style for writing the abstract, 13 journals note that this section must describe the paper and indicate its importance; the abstract is highlighted as a general overview of the paper, enabling its dissemination; it is especially important to choose appropriate keywords to make the text visible. The
guideline does not refer to technical principles to guarantee the formality of the abstract, but to aspects that promote the retrieval of articles in search engines. Thus, authors who submit texts in Wiley journals, for instance, are directed to a topic from the section Author Guidelines for explanations related to Search Engine Optimization (SEO) for their article, which includes an indication of essential aspects and keywords in the first two sentences of the abstract since only these sentences appear in the search engine results. Authors are advised to repeat these keywords three to six times. For an optimal choice of keywords, Google Trends and Google Adwords are suggested. It certainly represents an innovation for authors/researchers who are asked to act as marketing agents of their own work. As for the abstract, it is also important to highlight the innovation provided by Elsevier Publications, which recommends that authors of articles published in its journals should elaborate a graphical abstract, a single, concise, pictorial and visual summary of the main findings of the article. The graphical abstract will be displayed in online search result lists, the online contents list and the online article. Graphical abstracts allow readers to quickly understand the main message of the article. It is also intended to promote interdisciplinarity and help readers quickly identify articles relevant to their research interests.

In CS, 96.9% (n.31) of the sample journals provide guideline for writing abstracts. In this group, there are 27 journals that determine the number of words, ranging from 100 to 250 words with different guidelines: 2 indicate from 100 to 150 words, 11 indicate a maximum of 150 words, 3 do not permit more than 200 words, one points to a limit of 250 words, five consider 200 words necessary and three journals require 250 words, while another journal requires 150 words and the last one allows from 150 to 250 words. With respect to the structure of the abstract, four journals present considerable differences between original research documents and review articles. For original documents, it is necessary to describe the method and proceedings. For reviews, there is a different approach: first the primary objective of the review, the reasoning for the choice, the main outcomes and the review outcomes and conclusions must be indicated, including the implications for new research, applications or practice. One journal requires a structured abstract covering the main factual points of the article, as well as a statement of the research objective or problem, method, results and conclusions while another requires a structured abstract with the following sections: "Background", "Objective", "Method", "Outcome", "Discussion" and "Conclusion". Concerning the style of the article, seven journals include relevant advice: one indicates not to start with “In this article...”, but rather to provide a statement regarding the article's key points of interest; six Sage journals highlight the importance of the abstract because it has free access and it is where search engines seek information, allowing its retrieval by means of a user’s search. For this reason, its guideline recommends repeating three or four descriptive key sentences, but without becoming abusive because Google can detect such language; five journals highlight the need for not using
abbreviations, footnotes or references, considering that the abstract can be used independently of the text which it describes. Two journals consider the possibility of creating and providing a video abstract, opening new ways for this component of scientific articles.

3.4. Guideline for determining keywords

Indications regarding keywords are given in 70% (n. 21) of the LIS sample journals presented in the author guidelines section. There are thus 20 journals with rules about the number of keywords. This number can vary between a minimum of three and a maximum of 12 keywords. There are five journals indicating 5 to 10 keywords and two journals indicating up to 6. The other indications are different. There are two journals that consider using five keywords or short sentences. One journal is more specific recommending that keywords be used in the abstract as well as giving an indication of relevant keywords and synonyms. It is highlighted that keywords are not only important for SEO, but they are also used by abstracting and indexing services as a mechanism to tag research content. Only 23% (n. 7) of journals include guidelines about a controlled vocabulary for selecting keywords. In one case, the possibility of using a thesaurus is admitted, but without specifying any, and in another case the use of MeSH terms is mentioned. Two journals present a list of recommended keywords. Only one journal indicates the use of a controlled vocabulary provided by the submission system. In addition, authors are advised to avoid abbreviations, as only abbreviations firmly established in the field may be eligible.

In CS, indications about keywords are found in 87.5% (n. 28) of the sample journals. Accordingly, there are 27 journals that require a specific number of keywords, ranging from a minimum of three to a maximum of ten. Within this range, there is a great diversity of keywords among journals. Thus, only three journals have similarities requiring 10 keywords, two journals require 6 keywords, four journals require 3 to 4 keywords, four journals demand 3 to 5 keywords, two journals require up to 10 keywords and another two require from 5 to 6 keywords. As for the type of keywords, only one journal indicates that the plural and the use of multiple concepts should be avoided, and that the use of "and", "of" as well as abbreviations must be restricted.

4. Discussion and conclusions

Due to the importance of titles, abstracts and keywords for the organization of knowledge in journal articles, this study analyzed the editorial policies under the rules and recommendations provided to authors of texts in a sample of journals indexed in the Journal Citation Report in Library and Information Science (LIS) and Communication Science (CS) in 2016.

The study evidenced the scarcity of scientific literature on the subject and, empirically, found a tendency for the standardization of journals belonging to the same
editorial group, sharing general indications and criteria, in such a way that these indications do not depend so much on the thematic area, or the position in the ranking, as on the editorial group.

On the other hand, the study reveals that the journals have a different behavior in LIS and CS in regard to their editorial policies on titles, abstracts and keywords. While 84.3% of CS journals present a standard or style guide, only 22.6% of the LIS journals do likewise.

The precise and expressive writing of titles of articles can contribute to improving visibility and retrieval from information systems. Indications about the elaboration of titles appear in 71.9% of CS journals and in 56.7% of LIS journals, included in specific sections inside the “Guide for authors”, “Format guidelines”, “Manuscript submission” or “Manuscript requirements”. These are concise indications about the maximum number of words and, to a lesser extent, about more appropriate terms, or terms that should be avoided.

Abstracts of articles are a powerful tool to facilitate access to original texts, and they are the most important aspect in the standards and recommendations studied. 96.9% of CS journals and 76% of LIS journals provide specific guidelines for writing abstracts in their articles. These guidelines differ somewhat between research articles and review articles. They deal with the number of words and, to a lesser extent, with structure and style. The most innovative aspects are a reference made by the publisher Elsevier to a graphical abstract on its platform, projected so as to allow readers to gain a quick understanding of the main message of an article, while two journals in the CS sample refer to a video abstract, opening new ways for accessing articles.

Keywords are the object of guidelines in 87.5% of CS journals and 70% of LIS journals. These indications deal with the maximum number of words and, to a lesser extent, with issues concerning appropriate terms and controlled vocabulary.

The study reveals that guidelines on writing abstracts, titles and keywords have little presence in the editorial policies of CS and LIS journals. These issues are more related to protocols of editorial management procedures than to the intrinsic quality of documentary products associated with scientific organization and representation.

The study shows that, in a digital context, it is necessary to increase the focus on rules and recommendations provided to authors of articles to improve the representation and retrieval of articles. This exploratory study concludes that it is necessary to probe more deeply and expand this analysis into more theoretical and applied research on the essential processes of knowledge organization and representation in editorial policies. It is also important to elaborate proposals for standardization and style manuals with instructions for authors of scientific papers.
References


Abstract
This paper presents a project to develop an ontology for the health area, specifically in the context of information organization and retrieval for infectious and parasitic diseases in the thematic field of dengue and yellow fever. The Virtual Health Library in Infectious and Parasitic Diseases (BVS-DIP) received many expressive requests from its “Contact Us” tool – a communication channel between the BVS and its users. In view of this demand, the BVS saw the need to create instruments capable of improving the efficiency of the information retrieval process so as to avoid ambiguity and overcome difficulties in finding descriptors that could satisfactorily interpret users’ searches. The concepts that constitute the DeCS controlled vocabulary are organized in a hierarchical structure enabling searches to be made on more specific terms that belong to the same hierarchical structure. Due to its specificity, it does not cover terms for organisms according to taxonomies adopted by the classical works in the area nor of diseases according to the criteria of the World Health Organization (WHO) or by the research groups from Fiocruz institutes. Ontologies can improve information retrieval processes by organizing data content and data sources in a given domain, refer to the use of concept vocabulary, as well as clearing unambiguous relationships. Finally, it is possible to ensure that those participating in the ontology contribute to the use of the search domain’s vocabulary, thus improving the information search and retrieval processes, as well as the effectiveness of searches undertaken within the framework of BVS.

1. Introduction
Access to digital information has led to the emergence of new types of resources that have provoked a paradigm shift in relation to traditional existing supports (Le Coadic 2004). The new information technologies are creating “libraries without walls for pageless books” (Browning 1993). Better known as virtual libraries, these new forms and supports are redefining the current paradigms of information, communication and even the context of professional work within the field itself. The Virtual Library, considered an instrument to facilitate the rapid localization of information, has been driving momentum in several areas such as the field of health sciences. It is in this field that the Virtual Health Library in Infectious and Parasitic Diseases (BVS-DIP) stands out.

The BVS-DIP is a national collective construction project, coordinated by the Pan American Health Organization (PAHO) and the Latin American and Caribbean Center for Health Sciences Information (Bireme). It is guided by a National Advisory Committee constituted by representatives of institutions involved in the area of Parasitic Infectious Diseases, under the leadership of the Manguinhos Library in the Institute of Communication and Scientific and Technological Information in Health (ICICT) at the Oswaldo Cruz Foundation (Fiocruz). The main objective of the BVS-
DIP is to promote the cooperative and decentralized operation of the source network of scientific and technical information on infectious and parasitic diseases in Latin America and the Caribbean, in order to provide equitable access and democratization of scientific and technological information. Besides being a facilitating instrument, its basic principle is that scientific and technical information is a priority in health, both for the realization of an efficient and effective decision-making process, and for the formulation of public policies for management or research. The search process is supported by a controlled vocabulary called DeCS – Descriptors in Health Sciences – created by Bireme / PAHO and developed from MeSH – Medical Subject Headings – for indexing informational resources in the health area, and which aims to serve as a unique language in the indexing of scientific journals, books, annals of congresses, technical reports and other types of material available in the BVS.

In view of the significant level of requests received through the "Contact Us" tool of the BVS / Fiocruz, the need was identified to create instruments capable of improving the efficiency of the information retrieval process, so as to avoid ambiguity and overcome difficulties in finding descriptors that could satisfactorily interpret users’ searches.

The recovery process addresses concern about the organization of information available in the Web environment. This organization should allow the information to be retrieved with greater precision so that the user does not get lost in the searches.

Thus, a proposal emerged to develop an ontology project. It can be considered as a component of knowledge organization and representation, and a pilot aimed at promoting the retrieval of information from the “BVS” in the fields of dengue and yellow fever. The term ontology originates from the Greek "ontos", being, and "logos", word. In knowledge organization, it is defined as a "catalog of types of things" (Sowa 1999). According to Almeida and Bax (2003), Ontologies can improve information retrieval processes by organizing the data content and data sources in a given domain.

An ontology embodies an approximation of the real world (Guarino 1997) and refers to the use of concept vocabulary, as well as clear and unambiguous knowledge relations in the target search domain. According to Campos et al. (2006), these concepts should be stated based on characteristics that indicate their composition and their associations in relation to the application context. The use of shared concepts can provide paths for the interconnection of information resources, as their facts are identified. The use of these concepts and their representation in models can contribute to broadening the semantics of the connections. It is noteworthy that the use of models with distinct levels of expressiveness is a current activity in the history of mankind (Allemang and Hendler 2011).

Therefore, it is necessary to use concepts related to knowledge organization (KO), which seeks to analyze concepts and their characteristics, the position that each concept
occupies and its respective relations within a representational system (Bräscher and Café 2010).

KO enables the structuring and systematization of concepts. These can be organized according to their characteristics, ideally contemplating those aspects of concepts involved in the relationship between parent-child (inheritance) which can be reflected in the structuring of concepts into classes (Dahlberg 2006).

Thus, the planned objectives in developing a domain ontology project are to semantically analyze concepts, classes, inheritances, relations and axioms, and, in so doing, structure and systematize them. In addition, the project specifically aims to align with the DeCS vocabulary, as a means of structuring the concepts of dengue and yellow fever.

2. Organization of knowledge and the ontology

Several specialists in the information field have argued that information and communication technologies (ICT) should be incorporated into archives so as to streamline information, whether in large libraries or information and documentation centers. According to Lancaster (2004), the rise of information technology, especially with the development of the Internet, is reflected in the development of library services and systems. The library, as an information system, will continue to have significant roles to play, so that librarians should see these changes as an opportunity to improve services and products, and to gain new users and increased visibility by expanding into more areas of society.

In this way, libraries should endeavor to follow the progress of new information and communication technologies, thereby enabling appropriate organization of knowledge and information in all its many documentary forms as well as facilitating access and retrieval of information in a virtual environment.

KO involves analysis of concepts and relations in a knowledge domain (Hjørland 2003). It scientifically orders the structuring and systematization of concepts according to their characteristics, which can be defined as elements of object inheritance. The application of KO enables concepts and concept classes to be ordered by indicating the reference or content values of objects and topics, making use of instruments such as thesauri, taxonomies and ontologies (Dahlberg 2006).

According to Campos (2012), KO can be defined as the condition of knowledge representation for specific purposes. KO is based on the analysis of concepts and their characteristics to establish the position that each concept occupies in a given domain, as well as its relations with the other concepts that make up a conceptual system (Bräscher and Café 2008). Moreover, Café, Barros and Santos (2014) emphasize that "KO aims to generate the representation of a worldview, a part of this knowledge, for purposes of practical use in activities related to information organization, which applies
to the world of physical objects, of documents”. Miranda (2006) emphasizes that organization and representation are fundamental aspects in the information retrieval process.

It is in this context that we highlight ontologies as a tool to aid information retrieval. The ontology can be understood as a common and shared vocabulary, encompassing concepts and relations in the target domain, transmitted through computational representations that enable machine interpretation (Marcondes and Campos 2008).

The term originates from philosophy, emerging in the passage from the Middle Ages to the modern age. It refers to the study of being, its characteristics and its conditions of existence.

Sowa (1999) proposes that an ontology in knowledge organization is defined as a "catalog of types of things". Marcondes (2009) emphasizes that an ontology is an informational model that describes and represents a specific knowledge domain. Conceptualizing the relevant objects in this domain, their structure and inter-relations, such a model must represent a shared understanding in the user community. Marcondes (2009) goes on to note that, firstly, ontologies provide a common conceptual framework with which we can develop reusable and shared knowledge bases. Secondly, they facilitate the interoperability and the fusion of information that enables the creation of powerful and smarter computing applications. Hence, ontologies are being used in order to solve the problem of information retrieval on the web. Ontologies allow for improvements to data source indexing, through semantic representation and user searches performed by delimiting the context.

An ontology defines a common vocabulary for a community that needs to share information in a particular domain. It includes definitions of basic concepts in the domain and the relationships between them so that they are interpretable by machine. Developing an ontology is a complex activity that consumes a lot of time and resources, and also involves making a whole series of political, epistemological and conceptual decisions (Campos et al. 2007).

Some requirements for the development of ontologies are presented by Gruber (1996), namely: clarity and objectivity, completeness, coherence, extensibility, ontological commitment, ontological distinction, hierarchical diversification and standardization. Ontologies may still be able to provide interoperability support, allowing distinct users to use the same set of terms, supported by a standard interchange format.

Thus, among the types of ontologies identified in the literature, we find, with respect

---
1 In the original: “OC visa gerar a representação de uma visão de mundo, de uma parte desse conhecimento, para fins de utilização prática nas atividades relacionadas à organização da informação, que se refere ao mundo dos objetos físicos, dos documentos.”
to the degree of formalism: ontologies that are informal, whereby they can be expressed in natural language; semi-informal, where natural language is structured to some extent; semi-formal, where they are expressed in artificial language that is formally defined; and formal – when terms are defined using formal semantics. Regarding their function there are: high-level ontologies, which describe broader concepts and are applicable to any domain; domain ontologies – which describe concepts and vocabularies related to specific domains; task ontologies – which describe tasks or activities that can contribute to problem solving; and application ontologies – which describe concepts that depend on a specific domain (Morais and Ambrósio 2007). In addition, researchers can explore a foundation ontology, a further categorization resulting from the opportunities created by applied ontologies, which searches for usage in order to create patterns and anti-patterns to aid in the modeling and respective conceptualization of domain knowledge (Guizzardi 2014).

This study adopted the domain ontology as an appropriate instrument to apply to the issue of information retrieval in the BVS-DIP.

3. The issue of information retrieval at Fiocruz

The BVS’s underlying belief is the fact that scientific and technical information is a priority for health, whether for an efficient and effective decision-making process or for the formulation of public policies for management or research. In order to meet this commitment, the BVS adopts the following guiding principles: technical cooperation should be used to strengthen and expand the flow of scientific information; and equitable access to information should be promoted in its entirety to enable knowledge-based health decisions (Guzzo 2009).

Hence, the BVS may clearly be considered a facilitating tool for locating information quickly in the virtual environment, a channel of increasing use for the dissemination of information. The ease and speed with which information is made available and retrieved in the virtual environment has been the driving force in several areas, such as in the field of health sciences. According to Packer (2005), the BVS represents a radical expansion of previous models of information management and knowledge in health, bringing innovations and challenges with it.

Aimed at providing equitable access and to stimulate the use of updated information, the BVS-DIP was launched to the public on December 9, 2004. Its objective was to promote the cooperative and decentralized operation of the source network regarding scientific and technical information on infectious and parasitic diseases (Packer 2005). The theoretical context adopted by this study deals with infectious and parasitic diseases, accessible through BVS-DIP.

The implementation of the BVS-DIP was based on terms concerning infectious and parasitic diseases used by Fiocruz researchers in their research and also the terms that
are listed in TEMAN² (the Manguinhos Thesaurus). The structure of this thesaurus consists of terms and relations (equivalence, hierarchical, partitive and associative).

Another component used in the implementation of the BVS-DIP was DeCS. DeCS is a structured and trilingual vocabulary (Portuguese, English and Spanish) organized in a hierarchical structure allowing for research to be conducted using broader or more specific terms or all the terms belonging to the same structure. It is also used in the research and retrieval of concepts from the scientific literature in the Lilacs and MEDLINE bibliographical databases and other sources of information from the BVS.

4. Materials and methods

In view of the proposed objectives, this study is characterized as descriptive and exploratory research following a qualitative approach (Gil 2007). As for the procedure, it can be classified as bibliographical research and field study.

The bibliographical research for the theoretical background, the documental research for data collection and analysis, was conducted through the analysis of texts obtained in libraries, the CAPES Periodicals Portal, the BRAPCI Database of Periodicals in information science, digital repositories of universities and events annals in the area of librarianship and information science. Key words used were: ontology, knowledge organization, information retrieval, ontology engineering and methodology.

For the field study, distinct sources of information and instruments of representation available at Fiocruz were analyzed. The sources used for this paper in the OntoDIP domain were:

- Bibliographic databases: Infectious and Parasitic Diseases Database, Lilacs, Medline, Cidsaude, Repidisca;
- International Organizations: PAHO, WHOLIS, Popline, Pubmed, Web of Science;
- LIS-DIP – Localizador de Informação em Saúde em Doenças Infecciosas e Parasitárias (Health Information Locator in Infectious and Parasitic Diseases) – which contains the catalog of information sources on infectious and parasitic Diseases available on the Internet and selected according to quality criteria. It describes the content of these sources and provides links to them on the Internet;
- Health Legislation – LEYES: Basic Health Legislation in Latin America and the Caribbean;
- DIP Indicators – Dados Básicos – Brasil 2011, Dados Básicos – Brasil 2012, Superintendência de Controle de Endemias (Basic Data – Brazil 2011, Basic Data – Brazil 2012, Endemic Control Superintendence);

The following domain model represents the collection of instruments analyzed.

2 The Manguinhos Library, in 1986, developed the Manguinhos Thesaurus (TEMAN). It was constituted with the terms most related to the families, genera and orders in Experimental Medicine.
The process for the development of the ontology project was carried out through the following tasks: identification of concepts in the domain, definition of classes, properties and associations.

5. Results

Analysis of the concepts present in the sources used (Figure 1) enabled the proposition of more abstract classes, allowing the OntoDip ontology class hierarchy proposal (Figure 2), represented in the Protegé\(^3\) software.

---

\(^3\) Free software for ontology design. Developed by Stanford, it has two versions so far: Web Protegé and desktop. The desktop version (v. 5.2.0) was the one used in this research. Available at: https://protege.stanford.edu/.
6. Final comments

The development of an ontology involves many considerations, ranging from conceptual modeling, with the consensual registration of characteristics and relationships, to implementation in representation languages. Subsequently, the instances were created with appropriate data and relational values, allowing for the evaluation and testing of the ontology generated.

This work constitutes a first attempt to pioneer a new representation process by proposing an ontology model in dengue and yellow fever for the BVS-DIP. The construction of this instrument will favor information retrieval in virtual environments so that this information is applied in an appropriate way without problems of ambiguity, which is the purpose of this study.

Finally, it can be ensured that the elements in the ontology contribute to the use of the searched domain vocabulary and that implementation of the ontology allows for improvements in the search processes so that users can obtain information efficiently and effectively.

References


Abstract Knowledge organization (KO) can be seen as an emerging area since the scientific society which named it (ISKO) was founded in 1989. This paper presents the partial results of a terminological study of a qualitative nature which focuses on the set of expressions considered as belonging to KO and commonly applied in the literature within this area. It aims to organize the terminology used in the discourse of the KO community by combining two different and complementary viewpoints, namely Hjørland's rationalistic approach and Singh's historicist perspective. In this study, we propose a method based on literary warrant to identify, date, classify and place KO terminology in such a way that the product works as an area map. The methodology which will be followed begins with the selection of a corpus of reference works compiling the terminology of KO. One of the products is a list of all the terms recorded; dating of terms, year of coining and coiner; and a table of matching terms found in the sources. The main results arising from analysis of a random sample of ten terms will be submitted in advance in this paper. In conclusion, the fact that KO reference works currently belong to the domain (glossaries, dictionaries, classification systems, standards) provides ample indication that KO is gaining strength as an emergent new discipline which is in the process of stabilizing, defining and organizing its key notions.

1. Introduction Knowledge organization (KO) can be considered as an emerging area since the scientific society which named it (the International Society for Knowledge Organization-ISKO) was founded in 1989. However, this field of study has been developing since the middle of the 19th century, incorporating more or less autonomous or overlapping areas such as classification, subject analysis, information retrieval and indexing, all of which belong to LIS. It is clear that these and other fields, such as informatics, linguistics, technology, semantics and terminology, converge in terms of theory, methodology and applications.

Although it is understood that KO is an intersection field between these and other disciplines, LIS seems to be “the central discipline of KO” (Hjørland 2008, 86), where it originated and from where it projected itself in various different directions. For Smiraglia, KO “is critical for the proper functioning of the Science of Information. Without that which is learned in KO, information retrieval cannot work” (Smiraglia 2014, 3). If we become aware of the multidisciplinary composition of this research field, as well as the various epistemological approaches which try to apply KO within other disciplines, we will have a better understanding of the emergent and developing processes of its terminology. Several authors have often wondered what KO is (Dahlberg 1993; 2006; Hjørland 2008; 2013; Smiraglia 2014).

Narrowly speaking, “KO as a field of study is concerned with the nature and quality
of such knowledge organizing processes (KOP) as well as the knowledge organizing systems (KOS) used to organize documents, document representations, works and concepts” (Hjørland 2008, 86)”. This dichotomous division of the domain which could be seen as a rationalistic approach to KO, implicitly proposing a border between the terms which can be associated with processes such as classification and indexing, their premises and those linking KOS (system typologies, terms found in standards, guidelines for system construction, domain analysis, etc.).

A historicist approach is also admissible, such as that proposed by Singh (2001) to describe and explain the terminology belonging to Ranganathan's colon classification, in which the background is ordered in a linear sequence starting with the contributions made by Richardson (1860-1939) and Sayers (1881-1960). This perspective, based on the timeline of canonical works and authors, can provide a convincing explanation of the terminological evolution within the field.

This paper presents the partial results of a terminological study of a qualitative nature which focuses on the set of expressions considered as belonging to KO or commonly applied in the literature within this area. It aims to organize the terminology used in the discourse of the KO community by combining two different and complementary viewpoints, namely Hjørland's rationalistic approach and Singh's historicist perspective.

The research objective is to carry out a historical and semantic reconstruction of the KO terminology area, to be used as the basis for research on single topics as well as the development of diachronic, synchronic and comparative studies.

Terminological research offers some comparative advantages with regard to other approaches, because not only does it analyze specialized language commonly found in the construction of science and technology discourse, but also studies interaction with the semantic web as well as the multiple aspects related to educational processes. Specific research focuses on a single term or on a limited number of terms from a particular thematic area (Talaván Zanón 2016).

2. Research context

If we look at the historical development of KO terminology, we can see that initially, as concepts began to take shape in the literature, they were first recorded in prefaces and tutorials of classification systems, in certain canonical texts (i.e. Cutter 1876; Richardson 1901; Hulme 1911; Bliss 1929) and, non-specifically, in the groundbreaking dictionaries and vocabularies of LIS (Harrod 1938; Thompson 1943).

The steady flow of terminology increased greatly around the middle of the 20th century due to the development of thesauri, and the contributions made by Taube (1955), Sayers (1959), Lancaster (1972), and especially Ranganathan (1967). In order to develop his faceted theory and method, Ranganathan created his own terminological system, made up of some 200 terms.
This trend continued with publication of the first standards (i.e. ANSI 1974), an increasing number of journals and the regular production of specialized dictionaries in LIS (Landau 1958; Buonocore 1976; Clason and Salem 1992). With the advent of the Internet and ICTs, and after the establishment of KO as a disciplinary field, a classification scheme (Dahlberg 1999), a digital encyclopaedia (http://www.isko.org/cyclo/) and three dictionaries of KO (Wellisch 2000; Gnoli, Marino and Rosati 2006; Barité et al. 2015) were published.

It is important to highlight that for decades standards have been an important tool to organize field terminology since they usually include a glossary for ordering and defining terms that require an explanation in order to apply standards effectively.

The aim of this paper is to propose a method based on literary warrant (the data provided by the topics present in documents) for the identification, dating and classifying of KO terminology so that the product works as an area map. It is thus potentially helpful to use the literature produced over the last 150 years as the basis for a historical and conceptual reconstruction of the domain, and for the setting of KO terms. In this way, we seek to establish the traceability of each term from when it was first coined until its current usage.

3. Methodology

The methodology outlined below will be adopted in order to fulfil the following objectives:

1. Selection of a corpus of works in several languages, compiling – in whole or in part – the terminology of KO. The core of this corpus will be constituted by at least two dictionaries (Gnoli, Marino and Rosati 2006; Barité et al. 2015) and four glossaries specialized in KO (Dewey Decimal Classification 2011; BSI 2005; NISO 2010; ISO 2011). The corpus also includes: all available glossaries and dictionaries specialized in LIS, canonical works, articles and submissions focused on a particular topic and the collection of the College & Research Libraries journal, published since 1939, and significant because of its subject coverage over a long period of time.

2. The creation of a database supported by TemaTres software. A terminological record file was established with the following items: term in English, Spanish, Portuguese and Italian, synonym, generic term, specific term, associated term, corpus definitions, original definition, theory or trend, core or fringe term of KO, bibliographic references of documents with in-depth analysis of the term, links to documents available on the Internet, coining date and author, notes and dating.

3. Extraction of data from the corpus and record in terminological files. The procedures used to collect information are those routinely carried out in the methodology for specific term research (Célestin, Godbout and Vachon-L'Heureux 1984) given that they include a retrospective analysis of each term,
from its coining to the present day, establishing the scope of the concept and any changes or extensions of meaning which may have arisen.

Literary warrant is taken into account since it was established that the only source from which to select and extract terms should be documents of the area. The only terms selected are those which have been mentioned by at least three researchers in the field and which have been in use in the literature for at least ten years. Terms in use (i.e. “folksonomy”) as well as obsolete terms (i.e. “uniterm”) have been included in the database, provided there is sufficient literary warrant.

As we are aware that this project can only be developed in the long term, this paper provides an advance submission of only ten terms from a random sample.

4. Results

The following comments result from the literature review process:

1. Part of the literature - especially scientific articles or presentations submitted to congresses - is devoted to the analysis of single terms. Thus, several authors have presented in-depth work on the notions of “relevance” (Mizzaro 1997), “folksonomy” (Noruzi 2006), or “facet analysis” (La Barre 2010).

2. Another portion of the documents studied focus on a KO sub area (i.e. Lancaster 1972), or on terminology established by an author or a trend (Spiteri 1998; Singh 2000).

3. Many authors suggest “neoterms” to name what they consider new phenomena, objects, operations or concepts, but a high percentage do not achieve peer recognition, since they are not convinced of the originality or innovation of the proposal. As a consequence, a great number of terms are usually not used beyond the sphere of action of the coiner and his closest students. Such terms do not fulfil the methodological framework established in the project and are, therefore, not included in the main database.

As a summary presentation of the results, Table 1 shows a list in English of the ten selected terms with their equivalents in Spanish and Italian, the year the term was recorded in the sources and, as examples, some coining dates. A column of comments and notes is also included.

Due to the lack of space, the chronological list of sources will not be included here. Only the year of the term’s recording in a source will be mentioned in the “Years of record (sources)” column.
### Table 1: List of terms and equivalences

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
<th>Italian</th>
<th>Years of record (sources)</th>
<th>Coiner, coining date in KO</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aboutness</td>
<td>Tematicidad</td>
<td>Tematicidad</td>
<td>1999; 2015</td>
<td>Fairthorne 1969</td>
<td></td>
</tr>
<tr>
<td>Classaurus</td>
<td>Clasauro</td>
<td>Classauro</td>
<td>2006; 2015</td>
<td>Bhattacharyya 1982</td>
<td></td>
</tr>
<tr>
<td>Granularity</td>
<td>Granularidad</td>
<td>Granularità</td>
<td>1988; 2015; 2017</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 5. Conclusions

The kind of systematization proposed is based upon the idea that sciences require a certain uniformity of criteria and that the reliability of their classifications is reduced to a minimum so as to facilitate a greater consensus among specialists. This research can also contribute to the chronological reconstruction of KO terminology, visualizing...
the periods of inclusion of terms and their origin.

It is possible that the first terms of the area appeared in order to name concrete practices of subject representation in libraries, their tools and products (*i.e.* “classification system”, “subject heading”) with the subsequent incorporation of others, of an abstract nature, derived from reflection on this practice (*i.e.* “aboutness”).

As a thematic area with strong interdisciplinary links, KO has reached a certain degree of maturity, attracting researchers from other areas, with a sustained scientific production throughout the world, accounting for the ongoing problems of subject representation of knowledge in all information contexts. This process is seen when the area terminology receives in-depth analysis from historicist and semantic perspectives. The combination of both perspectives naturally allows for a significant number of terms to be associated with the main researchers and scientists of the area, responsible for their creation, introduction or extrapolation from other knowledge areas (Dewey, Ranganathan, Hjørland, Beghtol, Sayers, Chaumier, among others).

The fact that KO reference works currently belong to the domain (glossaries, dictionaries, classification systems, standards) provides ample indication that KO is gaining strength as an emergent new discipline which is in the process of stabilizing, defining and organizing its key notions.

References


BSI (2005). *Structured vocabularies for information retrieval: BS 8723-1*. [S. 1.]: BSI.


Hemalata Iyer, K.S. Raghavan

Medical ontology: Siddha System of Medicine

Abstract
Examines the range and nature of concepts and conceptual relations in the domain of Siddha medicine – a holistic system of medicine that originated in India and also practiced in other parts of the world. While his medical system has developed extensively and the consumer base grown widely, its representation in the Knowledge organization systems has been inadequate. Discusses Siddha ontology, a knowledge base built to support semantic information retrieval.

1. Introduction
In the trend of globalization and worldwide integration, information transfer and use, capturing and sharing knowledge assets for enhancement and sustainability has become the need of the day. Knowledge management and specifically knowledge organization systems (KOS) have been created to facilitate this process. KOS, includes knowledge structures such as taxonomies, subject headings, controlled vocabulary, and ontologies. In the medical domain, the Medical Subject Heading (MESH), used for indexing biomedical literature and the Unified Medical Language System (UMLS) are examples of KOS.

Despite the advances in healthcare technology in conventional medicine (Allopathy) the use of Complementary and Alternative Medicine (CAM) has increased considerably in the past years (Iyer and D’Ambrosio 2013). The World Health Organization estimates that approximately 80% of the world’s population relies on traditional systems of medicines for primary health care, defining these systems as those in which plants form the dominant component over other natural resources (Mukherjee and Wahile 2006). Siddha medicine is a holistic medicine that originated in India and is being practiced in many parts of the world.

Information systems have, for long, employed traditional knowledge organization systems (KOS) such as classification schemes, thesauri, etc. for knowledge representation. The range of conceptual relations these can represent is largely limited to hierarchical and associative relations. Knowledge available in many of the domains is so complex that effective retrieval and knowledge discovery require more powerful tools and techniques of knowledge representation than what traditional KOS can support. The use of ontologies to overcome the limitations of keywords-based Boolean search has been projected as one of the key requirements of Semantic Web. Ontologies can represent more complex relations in a more precise and specific manner (compared to mere RTs in Thesauri). More importantly, Ontologies support inferencing and thus facilitate improved retrieval. Ontologies are projected as knowledge bases which can
be exploited to support semantic searches.

The Siddha system of medicine has not been adequately represented in any traditional KOS. In this paper an attempt has been made to examine the range and nature of concepts and conceptual relations in the domain of Siddha medicine and model these in an ontology.

2. Siddha System of Medicine

This is one of the oldest systems of holistic medicine that emphasizes both prevention and treatment of diseases. The fundamental principle of Siddha, as of Ayurveda, is geared towards maintaining equilibrium of three humors, or the metabolic agents in the body, namely, Vata (air), Pitta (fire) and Kapha humor (earth and water). The normal ratio between the three humors is 4:2:1 respectively. A disequilibrium between the humors is the underlying cause of disease. Amongst other factors, stress, environment, comprising of climate, diet, physical activities trigger the imbalance. Healthy living consisting of a disciplined regimen along with proper diet and medicines are recommended to restore body’s equilibrium. The government of India regulates the medical practice, standardizing and testing drugs, supporting clinical investigation of herbs, establishing Siddha research centers and Siddha medical colleges.

Originally, Siddhars (founding fathers) codified their “findings on the characteristics of plants, metals, minerals and animal products and their knowledge of the properties of drugs, its purification, processing, fixing dosage, toxicity, antidote and clinical application, were preserved in the form of verses for the use of the posterity” (National Institute of Siddha). Subsequently, the oral literature was transcribed on palm leaf manuscripts that continue to serve as a repository of Siddha knowledge complementing the substantial publications, journals, books and digital information.

3. Diverse user categories

One of the authors (Iyer) conducted extensive interviews with the Siddha practitioners and educators in India regarding the principles and practice, communication patterns and knowledge structures. Their input informs the ontology that is reported in this paper. There are diverse categories of users seeking and utilizing information and the information sources range from palm-leaf manuscripts to digital sources.

a) Practitioners: Formally educated physicians in Siddha Medical Colleges
b) Medical College students, researchers and professors
c) Vaidyars: Practitioners who typically have inherited their knowledge from their forefathers /Gurus. They apprenticed with a teacher in a Gurukula system (residence with the teacher). The apprenticeship involved a highly disciplined way of life, learning, reciting Siddha verses, observation, and mentorship. In essence, this facilitates a complete and total immersion. The teacher gave
individual attention to each of the students that resulted in a strong teacher-student bond. The students not only to gain knowledge and skills but also attitudes, professionalism and considerable practical clinical experience. One of the aspects examined during the interviews by Iyer was whether the two seemingly disparate groups of vaidyars and the formally educated practitioners professionally interacted/ networked. The vaidyars by and large focused on the specific areas of expertise, often as that of their forefathers/ teachers. The professional schools tap this vast resource. The interesting aspect is that though a dichotomy exists, there is mutual respect and acknowledgement that allows collaboration and consultation. This is rather unique to this system. Interestingly, the national level research councils also maintain considerable connection with the vaidyars.

To sum up there are distinct categories of users that need to be served. Their backgrounds, education and language may be very different and so also their modes of communication and access to information. The use of technology for interaction is not as much as one would expect from professional groups.

4. Siddha concepts and categorization

The following sections presents the description and classification/ categorization of the various components of the domain that form the basis of this ontology:

a) Disease: During the onset of a disease a single humor gets aggravated from its state of equilibrium. Diseases are fundamentally classified according to the three humors of the body. Thus there are Vata correlated diseases, Pitta correlated ones and so on. Symptoms, etiological and other factors are also used to further categorize.

b) Treatment: Siddha system uses vegetables, animal and mineral products in the preparation of drugs. Examples of the categories of medicine are: Kudineer chur anam--decoction powder; Thailam– medicated oil; Parpam– prepared by the process of calcination; Legiyams and Rasayanams – Contains herbal powder; Chunnam – alkaline in nature.

Being a holistic system of medicine the treatment techniques also includes lifestyle modifications, massage, meditation, breathing techniques, and diet, as a regimen for healing. The major categories of techniques are: Yogasanam– yoga postures; Varmam– manipulating vital energy points; Tokkanum– massage technique; Unavu ozhukam– diet regulations.
5. Diet in Siddha

Regulating diet and correcting lifestyle activities is emphasized for preventing and curing diseases. Based on the body constitution and seasonal variations, diets are prescribed to increase or decrease the three dosha (humors). Broadly, grains (wheat, rice and millet), fruits (banana) and dairy products increase Kapha. Spices (pepper, chili) increase Pitta. Staple food and legumes increase Vata. The Siddha literature provides specific details of the kind of food that regulates and normalizes each of the doshas.

a) Spices are correlated with doshas. Spices are rich in antioxidants and low in calories. Siddha recommends the use of eight spices namely: turmeric, cumin seeds, pepper, cardamom, dried ginger, fenugreek, asafetida, and garlic. Specific spices that aid healing of specific ailments are also indicated. For example: Curcuma longa (Siddha name, Manjal) possesses anti-inflammatory properties and is effective for healing wounds, common cold etc. Fenugreek (Siddha name, Venthiyam) stimulates insulin release and decreases blood glucose level.

b) During treatment specific diets are prescribed based on the disease, potency of medicine, period of administering drugs, climatic/ environmental changes etc. For example compatible diet after medicated oil bath is: tender lentils, certain types of dried fruits, tender mango, gooseberry and certain type of leaves (Mutthulakshmi and Vijaykumar 2009).

c) The combination of spices and ingredients to use, the type of utensils for cooking, menu combinations that help alleviate the impact of individual doshas are recommended. Details of the number of meals, quantity, exercise regimen, seasons and climate also are factored in the choices of food, spices, herbs etc. The underlying purpose is to improve digestion, absorption, and to ensure that the value of nutrition is not diminished.

6. Information resources

There is voluminous literature in the field, consisting of books, journals, conference proceedings, CDs and palm leaf manuscripts that are housed in Siddha libraries. The Palm leaf manuscripts are central and unique to this domain. Codes, symbols and poetic verses are used to express ideas. Therefore reading, writing and deciphering the manuscripts is a very specialized skill and requires scholarship and training.

There are thousands of manuscripts held by Siddha families. Individual vaidyars prepare medicines in accordance with the instructions in their manuscript collection. The formulas are kept very private and guarded as a secret family legacy. Hence, certain vaidyars are reputed for their effective medicines for certain diseases. The Government of India has several initiatives to locate, collect, digitize and preserve this invaluable treasure.
7. Siddha ontology

This section briefly presents the Siddha ontology. Protégé version 5.2.0 is used for developing the ontology.

The Principal Classes and Object Properties:

The major classes (and subclasses) in the Siddha Ontology (see Figure 1 and 2) developed for the present study are:

- Disorders: the set of subclasses that includes various health problems, diseases affecting the human body and mind. Again, while many of the concepts are common to other systems of medicine, there are some that are unique to Siddha. The Siddha system of medicine is based on the premise that all health problems are directly related to tridoshas. Thus, disorders are broadly viewed as vata disorders, pitta disorders and kapha disorders.

- Symptoms: set of classes that includes both physical and psychological manifestations when there is disequilibrium in the tridoshas. Again, these could be classified into vata disorder symptoms, pitta disorder symptoms, and so on.

- Medicinal Plants: Medicines (drugs) in the Siddha system are mostly plant-based. In this ontology, for representing the subclasses of medicinal plants both the common names (mostly Tamil names) and botanical names are used. These plants also have therapeutic value.

- Siddha Medicines: Drugs made of herbs / plants etc. prescribed for treatment of various disorders.

- Therapy: The major curative and preventive measures prescribed in Siddha system.

- Institutions: that teach, practice and / or conduct research in Siddha system of medicine including hospitals.

- People: Siddhars (who developed the system of Medicine), Vaidyars, who practice the system as a family vocation, and authors of publications constitute this class.

- Adverse Effects: Some of the medicines when administered may have certain undesirable effects. For some of these classes a few instances have been added purely for the purposes of illustration.

- Documents: include works (classics by Siddhars), journals, scholarly papers, manuscripts, etc. One way of classification (subclasses) could be Texts and Media.
The Principal Object Properties that are used in the Ontology to relate individuals are shown in Figure 3:
Searching the ontology

It has already been mentioned that an ontology is a knowledge base and is built to support semantic information retrieval. An entire information corpus (not merely documents) can be fully represented in an ontology supported knowledge base. The search of such a knowledge base results in retrieving ontology entities. While this work is still in progress to be able to exploit the full capabilities of an ontology-driven knowledge base, the domain ontology populated with many classes, subclasses and instances is adequate for testing purposes. Some typical results using search queries involving regions of the ontology that are complete to a higher degree than others are presented below. The following figures (Figures 4 & 5) present the results of a couple of searches carried out for purposes of demonstration.

Query 1: Figure 4 presents the results for the query “Symptoms of Vata Disorder”.

Figure 3: The Object properties (only illustrative)
Figure 4: Query 1 – Symptoms of Vata Disorder

Query 2: Figure 5 presents the results for the query “Institutions of Siddha Medicine”
The value and utility of semantic retrieval as compared to keyword-based searching is clearly evident. A demonstration of the degree of enhancement in retrieval is dependent largely on the completeness and quality of the domain ontology. For example, if the knowledge base is complete with documents on prescribed therapy for treatment of a particular disorder in different systems of medicine, it is possible to query the system to retrieve values from structured data.

8. Conclusion

Since the original texts are written in Tamil, that too in the form of poetic verses, kural, nool, padal, it poses a language barrier to users. The treatises that are in English also tend to use the Tamil medical terminology. Hence it is important to use Tamil terms with English/ botanical/ scientific equivalents while representing concepts in the Siddha ontology. This will help to universalize the concepts, to preserve cultural sensitivity and aid better communication. This is true of names of diseases, spices, herbs, medicinal plants, food ingredients, medicines and treatment, all of which are expressed using Tamil vocabulary. In addition, employing annotations liberally to explain the abstract and unfamiliar concepts will help the users.

An ontology-driven knowledge base can perform complex queries on structured data and it is important to build such knowledge bases. An issue encountered in the present context is that some of the concepts in the domain are unique to the culture in which the system developed. Exact equivalent labels, for example, in English for some of the concepts may not exist. Further experimentation, larger knowledge base and corpus of documents are needed to improve the ontology. This work is in progress.

References


Carlos Henrique Juvêncio, Georgete Medleg Rodrigues

The Modern and its impact on models of information organization in Brazil: the decline of the National Library and the rise of the National Book Institute (1930-1954)

Abstract
This paper seeks to discuss how the National Library of Brazil lost ground on the national scene after the coup d'état led by Getúlio Vargas in 1930 and even more so after the founding of the National Book Institute (INL). It highlights the fact that there was a change in the models followed by Brazilian libraries with the European model being replaced by the North American. It shows that several of the functions that previously belonged to the National Library were incorporated into the INL when it was founded, which contributed to the contemporary decline of the National Library, both in relation to other libraries and in terms of the techniques used in its wide-ranging collections. It concludes that the INL served as a means of propaganda for the national culture as it supported the Brazilian publishing industry and encouraged the creation throughout the country of libraries promoting national knowledge.

1. Introduction
At the beginning of the 20th century, the Brazilian National Library under the directorship of Manoel Cicero Peregrino da Silva underwent several changes, including the inauguration of a new building and the publication of a new regulation, although the biggest change implemented in the institution was the application of Otlet’s precepts of documentation to the organization of its collection.

Undergoing a profound transformation, the National Library adopted documentation precepts, becoming the Brazilian hub of a worldwide network of information exchange. In addition, it assumed national leadership of initiatives in favor of universal knowledge and the organization and diffusion of the country’s intellectual production. Accordingly, it inaugurated a bibliography and documentation service, in the mold of Otlet and La Fontaine’s proposals, which set about trying to publish the National Bibliography. In addition, the National Library served as a consultation warehouse for the National Bibliographic Repertoire (RBU), an information service which operated as a kind of reference desk for users to locate information they wanted, whether it was in the Library or not.

Moreover, the International Exchange Service was expanded, and contact with various institutions around the globe became more intensive and profitable. While receiving several publications from abroad, the Library also sent out its own, thus serving to disseminate both Brazilian intellectual knowledge and propaganda about the ideals defended by the newly proclaimed Republic, which demonstrates Brazil’s potentialities and how the country sought to develop in the wake of republican ideals.

However, in 1930, a new regime was installed in Brazil, and the so-called First
Republic ended, initiating a period that subsequently saw the establishment of the Estado Novo (New State) dictatorship, with Getúlio Vargas as its ruler. In fact, the great characteristic of the period was the reorganization of the Brazilian state, with new institutions being founded and many others going through a process of restructuring. In informational terms, if until 1930 the National Library had a monopoly on matters concerning the organization of information in the country, dictating rules and serving as a mirror for the other institutions, with the coup of 1930 and the consequent establishment of the Estado Novo that hegemony was broken and the Library lost many of its functions to the newly created the National Book Institute (INL), which took the lead in library affairs and the organization of the national bibliography, for example.

This paper aims to highlight how the change of the axis of influence in the organization of information in Brazil occurred, shifting from Otletian Documentation, of eminently European origins, to North American Librarianship.

We will use bibliographical research, especially from articles published in the Public Service Review, where the Public Service Administrative Department (DASP) – a body superior to the INL – published several opinions on the organization of information in the country, and legislation regulating the functions of Brazilian public administration bodies. In addition, consulting the Institute’s archives, held by the National Library, also assists greatly, since it will help us to understand the social role of the institution.

It is clear that the change in the model of Brazilian government resulted in major transformations to the institutional political framework. This explains why the National Library – which maintained a prominent position in the context of the First Republic – was rejected by the new government, since new guidelines and symbols were being forged. As a result, the National Book Institute emerged with the goal of being an achievement of the leadership period of Getúlio Vargas, whose first government lasted from 1930 to 1945.

In addition, mention must be made of the distinct cultural context in evidence, namely the rise of the Modernist movement, which since 1922 had begun to dominate the Brazilian scene. Thus, the kind of modernism that impacted the national intellect was more connected with what was later called the American way of life and with issues focused on the formation of a Brazilian culture concerned with national ideals, and whose Italian fascist model seduced some of the rulers of the country, especially Vargas.

We thus highlight the role of the INL as a promoter of culture through books, creating and modernizing public libraries throughout the country, especially in the most backward states, according to the government at the time; as well as its publication of books about Brazil and, above all, Brazilians. Therefore, if on one hand the inspiring model was American, the culture to be valued was Brazilian. In general, if one of the forms of disseminating Brazilian culture was through the creation of libraries, the
model that served as an inspiration for their establishment and organization was the American one, focused on the organization of information, but also on the education of its users.

2. The National Library and Otletian documentation

The National Library played a prominent role in the development of knowledge organization in Brazil, gaining, for a start, a new and modern building, inspired by the most advanced to be found in Europe and the United States in the first decade of the 20th century. In addition, it assumed the leading role in national librarianship, a fact highlighted by the enrichment of its collection, which was already the largest in the country and now became even bigger. This was achieved, moreover, by being given regulatory powers which promoted its leadership in initiatives to contact similar institutions abroad.

There is no shortage of evidence indicating the National Library’s aim to further knowledge organization at that time, with many distinct services being set up: the Information Service – created to provide guidance on where and how its users could get the information they wanted; the International Exchange Service – one of the ways used to enrich the collection; the Librarianship course – which sought to train professionals to be able to deal with the documentary variety of its collection; the Bibliography and Documentation Service – the utmost example of the Otletian documentalist influence in the institution (Juvêncio 2016).

Symbolic of the age when Brazil was ruled by emperors, the Library – the second most expensive item in the compensation that Brazil had to pay to Portugal at the time of independence in 1822 – had continually seen its relevance diminished, whether from being housed in inadequate facilities, or from governmental disenfranchisement. Only after 1900 did it gain prominence in the national life of the capital (Juvêncio 2016; Schwarcz, Azevedo and Costa 2004).

If there is one thing that the Brazilian Republic knew how to do, it was the construction of symbols, especially from positivist ideals. The National Library was thus called to take its place as a symbol of national power and the modernization that was sweeping through the then capital, Rio de Janeiro, and serving as an example to other national institutions of how Brazil could prosper. In addition to its sumptuous building, a major legal apparatus was set up so that it could carry out its functions as a repository for national memory. In fact, the main highlight in this regard was the issue of a new and modern Law of Legal Deposit, which obliged all national printers to send a copy of each work published in the country to the institution. In contrast, the Library would be responsible for publishing the National Bibliography, in order to publicize what was produced in the country, be it by Brazilians themselves, or foreigners (Juvêncio 2016).
The last of the actions carried out by the Library, the National Bibliography launched its first issue in 1918, after sending an official to Belgium to study the techniques employed by Otlet and La Fontaine in the construction of the Universal Bibliographic Repertoire (Juvêncio 2016; Rayward 1975). However, following his death in 1921, the institution found itself without a substitute and the production of the bibliography took a hiatus until the 1930s.

3. The National Book Institute

Manoel Cícero Peregrino da Silva left the administration of the National Library in 1924, and in the following years the institution suffered from the absence of the greatest icon of its modernization. The 1930s began with a sudden change in the Brazilian government system. After a coup, Getúlio Vargas deposed then President Washington Luís, taking control of the country (Gomes 2013).

The Vargas Era, as it was known, lasted 15 years (1930-1945), and though he was elected for another term between 1950 and 1955, he ended up committing suicide in 1954, after a great scandal with opponents (Abreu and Beloch, 1984).

The Vargas Era is known for its attempt to organize Brazil’s state machinery and public welfare, with the foundation of various institutions and a series of social breakthroughs such as the right to vote and regulation of labor laws. One of its most prominent bodies was the Public Service Administrative Department (DASP), responsible for organizing and managing the entire national public administration. At that time, libraries, archives, museums and other similar institutions followed the dictates imposed by the DASP in the organization of their services.

The word in vogue in the period was “culture”. These institutions thus began to serve as instruments for the promotion of national culture, with the enlargement of their institutions and the establishment of monopolies in several areas, including the economic sector. It is no surprise that Vargas was adept at implementing the fascist ideals in vogue in Europe.

In terms of cultural promotion, a new institution was created in 1937, the National Book Institute (INL), whose main aim was to publish works about Brazil and Brazilians, as well as to encourage the creation of libraries in the country, actions that the National Library had fulfilled for years but was at that time denied.

In addition, it was its function: "to organize and publish the Brazilian Encyclopedia and the Dictionary of the National Language" – a way of promoting Portuguese as the mother tongue and the culture of the country; "To publish all sorts of rare or precious works that are of great interest to the national culture" - a means of reissuing the classic works of Brazilian writers; "To promote the changes necessary to increase, improve and reduce the cost of book publishing in the country as well as to facilitate the importation of foreign books" - that is, to promote the national book industry; and,
finally, "to encourage the organization and assist the maintenance of public libraries throughout the national territory" (Brasil 1937).

Empowered with the ideals of the modernist movement, which since the 1920s had preached of promoting national culture as a way of strengthening national identity, the INL was the portrait of the national progressive spirit. Gustavo Capanema, contemporary Minister of Education, stated that libraries were necessary because they were "centers of formation of the personality, of understanding the world, of self-education, in short, centers of culture" (Fundação Getúlio Vargas 1997).

Therefore, the underdeveloped states of the country were targeted in the promotion of national culture through the implementation of libraries. Its collection was obviously endowed with the national classics and a minimal structure to serve its users, most of whom were students (Revista do Serviço Público 1938).

In fact, an important publication regarding the organization of information in the country was the Revista do Serviço Público. Published by the DASP, this journal featured articles on how to organize libraries, museums and archives, all aimed at modernizing such Brazilian institutions, providing them with information on the organization of information in the most developed countries, now that they were among their number. North America provided the most significant influence on the discourse of Brazilian thinkers involved in information organization, who all considered the model to be followed was that of the US Public Library, and its ideal that libraries should provide effective assistance in public education since childhood.

Thus, the INL based its actions on the American model of public libraries, as opposed to the National Library, which since its foundation in 1810 had more affinity with European ideas of information organization. From the 1930s onwards, European documentalist ideology lost prestige in the Brazilian intellectual milieu, giving way to American ideology, including more widespread dissemination of the Dewey Classification (Juvêncio 2016; Revista do Serviço Público 1940).

Against this scenario, the National Library once again lost ground, and it was only in the reform to its regulation in 1944 that the North American ideology became present. Another factor that had a significant impact on the decline of the National was the regulation that assigned the INL the function of publishing the National Bibliography. In addition to a clash of competencies, this made both institutions compete against each other to determine who would first publish this work, demanding undue effort and causing unnecessary costs. In fact, it seems to us that the National Library would, nevertheless, have more to lose, having been relegated to the background, in competition with a more powerful and prestigious institution whose height could only result in losses (Revista do Serviço Público, 1940).
4. Final considerations

The National Library, during the period of the First Republic (1889-1930), played a prominent role in Brazilian public administration, and was used as a showcase for modern Brazil, which the newly proclaimed Republic sought to promote as a viable performer on the world stage.

However, since the coup d'état of 1930, models previously in vogue gave way to new approaches to management, including those for libraries, museums and archives. The model of inspiration was now the Brazilian modernist movement - which in the literary and artistic fields had a great deal of prominence, with its icons assuming public functions focused on national culture and education, most of their work naturally being displayed in libraries, archives and museums.

Indeed, even with the nationalist notion so in vogue, it is not surprising that the models used to organize such institutions, rather than purely national ones, were adaptations of US models. The Revista do Serviço Público, through its articles, lobbyed greatly for North American institutions to be the inspiration for those of Brazil. Thus, libraries developed their services along the same lines as their American counterparts.

In this context, the promotion of libraries was carried out by installing them right across the country, as a way of developing the culture and education of the local population, with the INL being the leader in this initiative.

It seems to us that the fact that the National Library was greatly identified with the deposed political regime meant that it played a secondary role in the new national organization, being surpassed by the National Book Institute as the new leader in the dissemination of culture and national science, as well as in the organization of the country's libraries.

References


The Onomasiological approach and the function of definitions in the elaboration of domain models in ontologies

Abstract
The capture of knowledge for the elaboration of domain models can take a semasiological approach from a linguistic perspective (from term to concept) or an onomasiological approach, from an extra linguistic perspective (from concept to term). In this article, we support the onomasiological approach devised by I. Dahlberg and we demonstrate the roles played by definitions in this activity. A definition, in this sense, is understood as the product of an agreement established between the modeler/ontologist and the domain expert; it reflects the intended model of representation of the modeled world.

Introduction
Domain models are obtained from processes aimed at high quality representation of knowledge of selected phenomena in a certain domain. In recent years, informational ontologies have been considered as formal systems to represent knowledge in a given domain, thus acting as a formal model to systematize the expression of the knowledge they manipulate. Since the 1990s, they have been conceptualized as an explicit specification of a conceptualization (Gruber 1993a). Formally, an ontology is a statement of a logical theory defining the set of representation primitives used to model a knowledge domain, such primitives being classes, attributes, and relations. Definitions explain this understanding, while formal axioms restrict interpretation (Santos 2010).

Domain models resulting from these conceptualization operations in informational ontologies can be defined as models that represent knowledge of an application area (subject area, mission area or problem area) and which should serve as a consensus representation (or model of reference) of a conceptualization shared by a given community. In order to conduct a conceptualization, we shall refer to I. Dahlberg’s concept theory, keeping in mind the referent, to present the procedure of how to capture domain knowledge via an extra linguistic approach. This method is established within an onomasiological approach.

In this way, we present the roles that definitions play in this process: they are at the same time elements that allow the capture of knowledge about a domain, but are also the products of an agreement established between the modeler/ontologist and the domain expert for elaboration of the intended representation model of the modeled world.

This article is organized as follows: To begin with (after this introduction), we present the types of definition in ontologies. We will then discuss procedures for
capturing domain knowledge within the context of elaborating domain models in ontologies. We go on to analyze semasiological and onomasiological procedures, before presenting the onomasiological method of I. Dahlberg’s concept theory, with respect to the referent. Lastly, we present the final considerations regarding this research.

Types of definition in ontologies

We can consider, in general, that the essential function of a definition is to be able to explicate a given definiendum (symbol to be defined) by means of definiens (a set of statements about what is required to be defined). Thus, in ontologies, elaboration of a definition presupposes determining the necessary and sufficient conditions that enable objects to be distinguished from one another in a given possible world.

Therefore, according to Marconi (1997), the meaning of a term (the semantic value of a sign) arises when competent speakers – for example, experts in a particular domain – consistently use a series of letters or symbols to communicate about something in the world. This is what is meant by the referent of the term – for example, objects, processes or attributes relevant to the activity in which the specialists are involved. The meaning is thus something like a mental representation (or concept) that experts have in mind when they use a term to refer to a type of thing in the world where their domain of action is delimited. In this case, the mental representation expressed by a term is composed of certain characteristics that are (at least in part) also presented in a definition. Hence, the term, the mental representation and the defining statement have the same referent; and, accordingly, definitions are elaborated with respect to the referent (Seppälä 2015).

A defining statement results from an understanding of the referent and the mental representation produced by the specialist and, as such, it should evince the consensus that has been agreed upon in a given domain. Thus, the explanation of a defining statement is realized in two ways in ontologies: by textual definition and by logical definition. Such classification is presented by Seppälä, Schreiber and Ruttenberg (2014), who state that ontologies have, on one hand, axioms that are part of a logical definition of terms and, on the other hand, definitions in natural language, that is, textual definitions. In an ontology, a textual definition is, ideally, a phrase that designates the object and its properties from a certain ontological commitment and whose function is to specify the intended meaning of the terms of the ontology in order to avoid ambiguities and human errors of interpretation. Logical definitions have axioms that restrict the intended meaning of a term by stating the necessary conditions for its use. They thus function in an analogous manner to the necessary conditions discussed in textual definitions and their function of disambiguation is also analogous. In addition, axioms enable the formation of taxonomic structures in ontologies and allow a machine to perform functions that will enable the automatic establishment of
inferences about the domain. Definitions result from a practice of authorship by the expert and modeler/ontologist. Such definitions, logical or textual, need not only contain elements that identify the closest genre and specific difference, but also others that can identify components, parts, function, purpose, cause, and so on. Definitions elaborated in this way are what Hegenberg (1974) called “operational definitions” – those in which the elements of the definition do not identify only the closest genre and specific difference – but also elements that seek to define certain operations wherein the concept would be applied.

In the field of terminological work, specifically related to the elaboration of definitions, several questions are presented related to the way definitions are constructed; in other words, whether they are in accordance with a linguistic or terminological perspective; whether it is possible to define something from the existence of the referent in the world or not (real or lexical definitions); whether there are differences between the concepts of description and definition; whether the definition is related to a communicative or referential function of the language, among other issues of fundamental importance with regard to the act of defining. Concerning the purpose of the present research, which we are now carrying out in ontologies, we will focus on the prime issue which is to call into question the linguistic and terminological perspectives for the capture of domain knowledge and which consequently relates to the elaboration of definitions, specifically to a given semasiological or ontological form of definition.

**The elaboration of domain models and knowledge capture procedures**

Domain models are obtained from processes that aim to represent knowledge of high quality phenomena selected in a certain domain.

In the elaboration of these models, methodological procedures are used for domain modeling, in other words, those related to conceptual modeling and those related to ontological modeling, where both approaches complement each other (Villela, Oliveira and Braga 2004, p. 243).

We also consider that both approaches complement each other. However, our experience has shown that, in the capture of knowledge for the elaboration of domain models in ontologies, ontological modeling, which arises from the extra linguistic, seems to us to be more appropriate, as we shall later argue (Campos 2010, Campos and Rocha 2017 and Gomes, Campos and Guimarães 2010).

Informational ontologies can thus be considered, as the study of the basic characteristics of a given reality. Aiming to explain the way in which these basic characteristics may be appropriated by the ontologist, we put forward two procedures that have been used to capture knowledge about a given domain, the semasiological approach, which takes a linguistic perspective and the onomasiological approach, which is from an extra linguistic perspective.
The semasiological approach considers the term as a lexical unit employed within the scope of a domain of specialty. Thus, the semasiological approach starts from the text and the possibility of extracting these terms from texts, since followers of this approach consider that the activity of surveying the terminological *corpus* to devise models presupposes the use of textual analysis to determine this *corpus*. Concerning ontologies, the adoption of this approach, in the majority of cases, with regard to the capture of knowledge, reflects a procedure involving analysis of electronic texts, starting from natural language and written text in its various formats (Santos 2010).

On the other hand, the onomasiological approach essentially considers that a term is reached through extra linguistic means, in other words, from an understanding of the referent within the scope of a field specialty, such that the term would be the result of an act of connotation with respect to the referent; in short, the term denotes the concept. Hence, while the concept is a mental representation, it precedes the linguistic form or any other symbol, since the relations between concepts are established due to extra linguistic parameters, from identification of the referent’s properties. In the field of informational ontologies, such an approach has been used in the stage of surveying domain knowledge, by means of conducting interviews with domain specialists. This procedure in the elaboration of ontologies has been called domain knowledge capture (Ceusters and Smith 2010, Fernandes, Guizzardi R. and Guizzardi G. 2011 and Tairas, Mernik and Gray 2008).

For the elaboration of domain models in ontologies, we consider defending the onomasiological approach. However, we will present a form of procedure on how to reach the extra linguistic that does not initially start with the expert, but with the establishment of queries about the referent within a domain. Such inquiries place the modeler in a position of researcher of the domain and not as a receiver of the expert's knowledge or even of his textual productions, as we shall attempt to demonstrate.

In order to realize a conceptualization, we will refer to Dahlberg’s concept theory, which is focused on the referent, to present the procedure of how to capture knowledge about the domain via an extra linguistic approach.

**Concept Theory and capture of domain knowledge: the onomasiological approach**

Dahlberg (1981) affirms that the understanding of something in a given field of specialty presupposes knowledge of three elements: a) its referent, that is, the concrete and abstract objects that meet in an extra linguistic space; b) the characteristics and properties of these referents, which are appropriate propositions stated within a domain with respect to this referent; and c) the verbal expression used and the that which promotes communication between a certain group of speakers. This is thus set within a conceptual triangle.

Concept theory (Dahlberg 1978a) provides a method for analyzing a reference item in a given domain, providing what we call here the capture of domain knowledge. This
method is established within an onomasiological context, because part of the analysis is of an extra linguistic element (the referent), about which questions are established in order to allow a certain systematization of this referent in a given domain. Thus, the predication that occurs with respect to the referent is not made in a random way; Dahlberg’s theory tells us a modus operandi of how the assertions about the referent in a given domain should be performed, as we shall see below. Here it is still necessary to emphasize the systematic aspect of concept theory; namely, that it allows the elaboration of definitions that are interconnected forming a cohesive whole which, in this case, represents the knowledge within a domain of specialty. That is why in Dahlberg’s studies she affirms that the concepts present themselves in relation to each other, forming a system of concepts.

Thus, her modus operandi begins with the identification of the referent’s properties. Accordingly, through the formulation of statements about the properties of referents, the characteristics of the respective concepts are obtained. For Dahlberg, the act of defining is a construction of selected aspects of the referent performed by the "look" of a group that acquiesces about its understanding in a given space of knowledge. In this way, these elements that form the concept – the characteristics – are properties of the referents (concrete or abstract objects) which, at the level of the concept, also become characteristics of the concept. The characteristics are, therefore, representations that are established from an "agreement" about the properties of the referents. Hence, 'concept', for Dahlberg, represents a unit of knowledge about the agreement signed within a knowledge domain.

Characteristics are also concepts and are used to classify and define other concepts. For example, one of the characteristics of the concept 'Resin-based varnish' is 'resin' and 'resin' is also a concept. If we do not know what resin is, we cannot correctly understand what a 'resin-based varnish' is. In this sense, the concept is never established in an isolated form, but always in relation to others. It establishes itself through comparison with other concepts, when similarities and differences are identified (through analysis and comparison of characteristics), which leads to the collection of concepts, related in a varied way. For example, there are similarities among the concepts 'fungicide', 'herbicide' and 'insecticide'. They have the feature of 'killing' in common. What differentiates them from each other is the object that suffers the action of the named substance: the fungus, the weed, the insect, respectively.

Identifying the characteristics of concepts is of fundamental importance for the understanding of the concept in a given domain, and consequently allows its classification or, in other words, its assortment within this domain. The characteristics are elements of the definition and reveal the connections existing between the concept in question and others through their relations. Dahlberg (1978b) classifies characteristics into essential and accidental. Essential characteristics are considered as the necessary and sufficient elements, in a given space of knowledge, to determine what
type a concept is, in other words, to determine its identity. Yet, accidental characteristics are optional for a certain type of concept, that is, they are not necessary or constant. Essential characteristics allow for assortment of concepts in the domain, functioning to structure domain models, such as taxonomies elaborated for the structuring of domain knowledge in ontologies. Because of this, she affirms that essential characteristics define; and accidental characteristics describe (Dahlberg 1978b). For her, the definition is ordered, so as to enable classification of the object in the domain. Hence, the first element of the definition must reveal the approximate genre.

A definition thus reveals an assortment of concepts from an ontological commitment (signed agreement) established in a given domain; essentially, it is the consequence of this assortment (classification).

As stated earlier, in this way, concepts define each other. In this process, certain relationships that provide the basis for their assortment are in evidence. Whenever there are two identical, opposing, similar, or other characteristics among two concepts, one can consider that there are relations between them. Relationships are classified as being quantitative and qualitative (Dahlberg 2014).

Quantitative ones involve a comparison of two concepts from a strictly formal point of view, including the following relationships: identity, inclusion, intersection and disjunction. Qualitative relationships consider formal and material aspects, and can be classified as formal/categorical; material/paradigmatic; and functional/syntagmatic. This network of relationships constitutes the system of concepts (Campos 2001, p. 104).

In the procedures for devising domain models, categorical relations are fundamental, because they allow a preliminary structuring of the concept within a domain. As Dahlberg (1978a, p. 34) affirms "categories provide at the same time skeletons, bones and tendons to structure all the knowledge”.

The categories presented by Dahlberg are metaconcepts that allow the modeler to reach any domain and succeed in establishing an initial alignment. They enable identification of referents within a domain. At the level of representation, they are highly inclusive classes in a domain, above which no concept can be placed. In her recent work, Dahlberg (2014) provides the following classification for categories, comparing the categories of Aristotle and Ranganthan: Entities (principles, material objects, immaterial objects) – which in Aristotle would be substances and, in Ranganthan, personality; qualities (qualities, quantities, relation) – in Aristotle distinguished according to categories of quality, quantity and relation, while, in Ranganthan, represented by the category of matter; activities – which in Aristotle are actions, processes and states, and, in Ranganthan, denominated as energy; lastly, dimension – which in Aristotle is considered according to the categories of place, time and condition, while identified in Ranganthan by the categories of space and time.

Categories are used for the ordering of concepts, but at the extra linguistic level of
the referent, they are also used to ask questions about the domain: whether in a certain domain there are principles, material or immaterial objects, qualities, and so on.

The onomasiological approach does not begin with a survey of a terminological corpus, but with identification of referents that can lodge in the categories - which function here as meta-representations of the domain. This way of entering the domain leads us, in the case of ontologies, in which it is our wish that possible knowledge can be explained, to be able to capture knowledge aided not only by the expert. This is a process of comings and goings, of analysis and synthesis. Thus, when the identification of which category the referent belongs to occurs, it is possible to determine relationships (logical, part - whole, functional) within these categories, in order to construct a domain model for ontologies. The classification of these concepts within the domain will then allow us to determine their enunciation by means of a definition that mirrors the established agreement.

Conclusions

We consider that the onomasiological approach, registered in a form of knowledge capture, enters the domain through understanding of the objects of knowledge within that domain; in other words, part of the extra linguistic allows for the formulation of a methodology that is potentially most appropriate for elaborating domain models in ontologies, since it enables a consensual understanding of how referents relate in a closed world.

References


OntoM4IS: ontology reuse method for Information Science

Abstract
Ontologies are a kind of knowledge organisation system (KOS) widely used in many contexts and for different purposes. When a new ontology is proposed for development, knowledge specialists will first decide whether it is possible to adapt existing ontologies. Our intention, with this article, is to propose a new method called OntoM4IS (Ontology Reuse Method for Information Science). The main idea is to reuse the content of ontologies, specifically their classes and properties, to provide the information science community with knowledge organization techniques. We will adopt a hybrid approach, combining the best features of each methodology. The process will be supported by design science research (DSR). We will also use qualitative research techniques, such as interviews and focus groups to validate our proposal. Moreover, as a research strategy we will follow the bibliographic research technique of published materials.

Introduction
Ontologies are commonly classified as a kind of knowledge organization system (KOS) such as classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabularies. Ontologies are widely adopted as KOS due to their properties for expressing knowledge (reality perceived by humans), promoting interoperability between human and machine agents.

The term ontology was originally associated to philosophy, but it is widely used in other scientific areas such as computer science (CS) and information science (IS). CS and IS borrowed the term ontology in recent years, especially with the exponential growth of Web, as a means of modelling reality and further representing it in a format (e.g. OWL – Web Ontology Language). Ontology has its roots in Aristotle, but in our field the boost came with the emergence of the semantic web, with its aims of promoting interoperability and the sharing (and reusing) of knowledge between agents.

The term ‘ontology’ (or ontologia) was coined in 1613, independently, by two philosophers, Rudolf Göckel (Goclenius), in his Lexicon philosophicum and Jacob Lorhard (Lorhardus), in his Theatrum philosophicum. Its first occurrence in English as recorded by the OED (Oxford English Dictionary) appears in Bailey’s dictionary of 1721, which defines ontology as ‘an Account of being in the Abstract’ (Smith 2003).

In this paper we address an important ontology design principle which is the reuse of contents from other ontologies. According to Ochs et al. (2017), reusing content (e.g., classes and properties) of reliable quality can save an ontology author significant time and effort. The same authors state that the reuse of content allows for a consistent

---

1 This article is funded by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil, under the PEC-PG Program
representation of a domain among all ontologies that reuse the same content. Support for reusing ontology content is included as part of the *Web Ontology Language* (OWL) specification (*i.e.*, using `owl: imports` axioms).

It is commonly accepted that reuse is important not only for the aspects mentioned but as a way of giving value to the work and effort of others, so-called credibility.

The authors Scherp, Saathoff, Franz and Staab (2011), state: "When designing an ontology, it is desirable to use a solid and sound modeling basis". Thus, our method recommends the use of a foundational ontology (also known as an upper level ontology) such as DOLCE (*Descriptive Ontology for Linguistic and Cognitive Engineering*), GFO (*General Formal Ontology*), or BFO (*Basic Formal Ontology*).

Foundational ontologies provide a high-level, abstract vocabulary of concepts and relations that are likely to be used in current and future application domains (Scherp *et al.* 2011).

1. Problem

The organization of knowledge is a crucial task and the use of ontologies is fundamental to fulfil this purpose. The literature review suggests a lack of methods for construction of ontologies in the IS field. However, some efforts can be identified, such as the *OntoForInfoScience* (Mendonça 2015), a methodological approach with concurrent support of three methods: *NeOn*, *Method 101* and *Methontology*. This methodology addresses an important yet unexplored aspect, namely reuse. Our object of analysis is to investigate the reuse of ontology content, focusing on classes and properties.

2. Objectives

The aim of this paper is to present *OntoM4IS*. The acronym OntoM4IS, stands for Ontology Reuse Method for Information Science. The main idea is to reuse ontology content, specifically classes and properties. Moreover, it is viewed as a contribution to knowledge organization within the IS field, providing a method to help researchers in matters of knowledge representation with the support of well-established methodologies.

3. Methodology

We will adopt a mixed approach, combining the best features of qualitative and quantitative methodologies. The process will be supported by design science research (DSR) (Hevner 2007), which essentially takes a pragmatic view, that is, a philosophical perspective which is problem-centered and based on the real world. We will also use qualitative research techniques, such as focus groups to validate some aspects of our proposal. Furthermore, as a research strategy, the study will follow the bibliographic research technique of published materials.
The DSR process, according to Hevner (2007), has three cycles (see figure 1): *Relevance Cycle*, *Design Cycle* and *Rigor Cycle*. These three cycles occur in three Spaces:

- **Environment** – this space is the context of the research, providing the requirements for the development of artifacts².
- **Design Science Research** – this is the core space, where the intersection of the three cycles occurs. This design cycle is responsible for the development of artifacts where the researcher builds and evaluates the design artifacts and research processes.
- **Knowledge Base** – this is seen as a support for the rigor cycle, providing the scientific foundations, experience, and expertise that supports the research.

![Diagram of Design Science methodology cycles](Hevner2007)

Gregor and Hevner (2013) proposed a framework to characterize and position DSR projects regarding the contribution of knowledge (see figure 2).

---

² The term *artifact* is used to refer to a thing that has, or can be transformed into, a material existence as an artificially made object (*e.g.*, model, instantiation) or process (*e.g.*, method, software). (Gregor and Hevner 2013, p. 341).
A project can be evaluated according to the maturity of the problem and the maturity of the solution. The authors define two axes and four quadrants. The maturity of the problem lies on the axis of the abscissa, and the maturity of the solution lies on the axis of the ordinates. As for the quadrants, they are the following: Routine Design, Improvement, Exaptation and Inventions.

The high maturity quadrant in both measurements (lower left corner), which Gregor and Hevner (2013) call the Routine Design quadrant, concerns applying known solutions to known problems. This quadrant makes no contribution to the knowledge base and does not, in fact, present any research opportunity since it constitutes existing knowledge that applies to routine situations (Malta 2014).

Gregor and Hevner (2013) call the next quadrant, the Improvement quadrant. This is a quadrant of high maturity in knowledge of the problem and low maturity in knowledge of the solution; it is about developing new solutions to old problems. This quadrant is an opportunity for research and brings possible contributions to the knowledge base (Malta 2014).

What the authors call the Exaptation quadrant is the quadrant of low maturity in knowledge of the problem and high maturity in knowledge of the solution; it is a question of extending known solutions to new problems. This quadrant is an opportunity for research and brings possible contributions to the knowledge base (Malta 2014).
Finally, in the upper right corner, the *Inventions* quadrant is the quadrant of low maturity in both dimensions; it is for inventing new solutions to new problems. This quadrant is a research opportunity and brings contributions to the knowledge base. It is a quadrant where there is a real start from nowhere; however, inventions are rare and inventors even rarer (Malta 2014).

Our method (OntoM4IS), in terms of knowledge contribution, can be positioned in the Improvement quadrant. OntoM4IS is based on contributions from information science and semantic web technologies (*e.g.* ontology development methodologies), which translates into an attempt to improve a known problem, by developing a new solution.

4. Related work

Reviewing the current state of the art has shown that proposed methods for the construction of ontologies are still incipient in the IS area. One research project has caught our attention, the OntoForInfoScience methodology, which is inspired by the upper level (and formal) ontology, BFO (*Basic Formal Ontology*). OntoForInfoScience is an ontology construction methodology based on three methodologies, NeOn, Methontology and Method 101. Reuse is a task that is emphasized and highly recommended in the proposal. This methodology has already been used and applied in the creation of two domain ontologies, Hemonto in the field of biomedicine and Ontolegis in the legal area.

Many studies have put emphasis on reuse. In fact, ‘Ontologies can be reused in several ways: they sometimes result in the creation of an independent ontology from the concepts of others (which can be extended and adapted), and in other situations they preserve the original ontologies’ (Campos *et al.* 2013).

5. Expected results

This paper will present a work in progress outlining the whole process starting with the literature review and further contributions, embodied in a method for reuse of ontologies in the information science field.

References


Robert D. Montoya

Parsimony in Biological and Colon Classifications

Abstract
The concept of parsimony has deep roots within a number of disciplinary discourses, including within philosophy, the history of science, biological classification, and Information Science. Parsimony as a concept is most often ascribed to William of Ockham (1287-1347), a Franciscan monk who wrote widely in the domains of logic, metaphysics, and theology (Wheeler 2012, p. 68). Ockham’s writings articulated an approach that helped one choose between competing theoretical models that came to be known as ‘Ockham’s razor.’ Ockham’s razor, in its most general sense, is “a theory that postulates fewer entities, processes, or causes is better than a theory that postulates more, so long as the simpler theory is compatible with what we observe” (Sober 2015, p. 2). On a practical level, if faced with a certain natural phenomenon, the most likely explanatory approach is to defer to the hypothesis that maintains a more economical series of causal events. Such assumptions about parsimony continue to be held in the domains of biological taxonomy and evolutionary biology.

Despite its origin and use in the domain of scientific theory production, parsimony is also implicitly (if not explicitly) used in the creation of taxonomies and classification within Information Science. We construct classification systems that should concisely represent and describe documents, as well as systems that must balance exhaustivity with enumerative economy. Additionally, parsimony is invoked when we define classes of things using a certain number of descriptive characteristics—too many characteristics and our schedules and notation become unwieldy, too few and we lose descriptive and exploitative power (Wilson 1968). Taking an interdisciplinary approach, then, this paper explores these issues and discusses how the principle of parsimony functions as an approach to producing and maintaining conceptually-efficient and optimally-coherent classification systems.

Identifying these modes of economization are essential, especially in light of the need to diversity our classification schemes and embrace multi-cultural and multi-linguistic approaches to organizing and describing our world. Calls for more pluralistic notions of classification, most strongly perhaps by Jens-Erik Mai (2011),1 are in conflict with many of the principles that underlie a parsimonious approach to classification system construction. We need to specifically address how our systems are simplified and economized to maximize their systematic efficiency, and to define how such efficiency is one way in which we limit potential uses of our classification systems. As Mai asserts, embracing a pluralistic view of classification acknowledges that “there are several conflicting but true descriptions of the world” (2011, p. 723). Knowledge is contingent and dependent on individual and cultural perspectives, so better understanding how our systems economize knowledge can help us embrace more contingent models of knowledge production and, ultimately, knowledge organization.

Methods and conclusions
This paper is a theoretical examination of parsimony as an integral aspect of classification production. Consequently, this analysis is guided by one basic question: How does parsimony function as a component of building classification systems both in defining classes and in building economical taxonomic models?

---

1 See also (Szostak 2015).
To answer this question, I take a broad approach to taxonomy and include a discussion of both biological classifications and traditional documentary/library classification systems, with a particular emphasis on S.R. Ranganathan’s Colon Classification.

Within the biological sciences, parsimony is a deep-seeded methodological approach to taxonomy construction. Oftentimes, taxonomies created within biological taxonomic spaces take an epistemological stance regarding how knowledge of the natural world should be constructed. The delineation of a “tree of life”—especially in evolutionary and phylogenetic taxonomic approaches—depends upon a series of statistical assumptions that balance the measurement of character traits changes over time. For example, it is often the case that an evolutionary tree that depicts evolutionary change with the minimum number of trait changes wins out as the preferred model—the assumption here is that the natural world defers to the path of least resistance in terms of such change, and thus our classification should as well. The talk examines the nuances of this argument, as well as exceptions to this approach (which undoubtedly exist in scientific practice).

Parsimony is also a concept that is invoked in the approach to classification in Information Science, especially in terms of how we manage and create economical and balanced classification systems. Ranganathan, for example, extolled the virtues of parsimony as a "normative principle" in the Colon Classification (1965) to economize the application of facets, the number of digits in class numbers, and the length of schedules. Ranganathan defines parsimony as such: “The principle that between two or more possible alternative rules bearing on a particular phenomenon, the one, leading to the overall economy of man-power, material, money and time considered together with proper weightage is to be preferred” (Ranganathan 1965, p. 63). For Ranganathan, parsimony is invoked not only in relation to the process of building of classifications, but also in terms of how much time and resources its takes to produce them – which have direct impacts on how usable a system is for an end user. The Colon Classification, then, as a system, is built upon the concept of parsimony.

The paper identifies four specific areas in which parsimony operates within classification systems:

1. The characteristics used to define a class
2. Intention and granularity of classes
3. Economizing temporal relationships
4. Controlling notation and expression

The paper concludes that the concept of economy and parsimony are key characteristics of the modernist approach to classification (Mai 2011), and that
alleviating the pressures for such economization are essential if we are to create more pluralistic systems. A call for more attention to this concept within the discipline is articulated.

References
Luca Giusti

The Penumbra-line: Ranganathan’s journeys and the genesis of the APUPA pattern

Abstract

This paper studies the background that may have influenced how the APUPA pattern was defined. It reveals how Ranganathan developed relationships with protagonists in the pragmatist movement and companies looking for answers to the problem of information overload. This work thus refines our understanding of the APUPA pattern, allowing it to be placed among the archetypes of new digital media, besides offering ideas for renewing information retrieval.

1. Classification and Communication

S.R. Ranganathan’s book, Classification and Communication, highlights how “container modularity”, which characterizes media like books, complicates their perceptibility (Ranganathan 1951, sect. 22-27 and 31).

Enlosing communication in containers guarantees the best transmissibility and storability, but, at the same time, requires a decoding activity, thus creating conditions for a paradoxical risk:

The quantitative strengthening of the channel transmits a large amount of information that, however, is not effectively usable; this ends up overloading the capacity for human processing (Ranganathan 1951a).

Ranganathan adheres to John Dewey’s analysis that the solution to overload risk is in large part a matter of depth; how deep we are willing to penetrate the bottomless pit of information:

Sometimes slowness and depth of response are intimately connected. Time is required in order to digest impressions, and translate them into substantial ideas. “Brightness” may be but a flash in the pan (Dewey 1910, 37).

Ranganathan’s proposal is based on an acute analysis of the articulation of user interest at different levels:

The boundary-line between irrelevant and relevant records is not sharp and clear-cut. Various degrees of intensity are possible…A totally or intimately relevant record may be called “Umbral Record” A partially irrelevant. “Penumbral Record” … A totally irrelevant… “Alien Record.” (Ranganathan 1951a).

He imagines that a renewed classification can produce the decisive ”transformation” from an indistinct flow to a deeply personalized one; it can do so by arranging records in an APUPA order:
should be arranged in a helpful order – *i.e.*, in the order Alien, Penumbral, Umbral, Penumbral and Alien so that all the most relevant records are in the centre and the others records stand fanned out on either side of the centre in decreasing order of their relevance (sect. 321).

![Figure 1: APUPA is often imagined as a bell curve](image)

The title of section, “Classification will have a future as a means of communication” (1951a, sec. 325), summarizes this vision: classification activity as a medium.

Classification is thus a transformation of thought. This change in arrangement from accession or random order to APUPA order is itself a subsidiary transformation. It is the first use of classification in the process of communication. The transformer is the classifier. The means is the Classification Scheme. (sect. 3212)

The communication flow of the new medium is schematized in Figure 2, recalling Shannon and Weaver’s model of communication.

![Figure 2: Ranganathan’s communication model](image)

We will now study the development of the issues which seem to have played a role in the genesis of the APUPA pattern.

2. Education in an accelerating society: Dewey

In 1920 John Dewey visited Madras and met the secretaries of the *Madras Teachers’ Guild*; Ranganathan was among them, working as a young math teacher.
In the early 20th-century, India experienced strong productive growth; accordingly, Ranganathan reported a very substantial list of new media reaching Indian students (1973 para. CB2).

In *How We Think*, Dewey (1910) repeatedly highlighted the risk of “overloaded curricula.” The book “New Education and School Libraries”, first published in 1942, is the clearest proof of the importance Ranganathan gave to Dewey’s new education ever since *School and Society* (Dewey 1899). It is a work that Ranganathan calls an “epoch-making book”, from which he reproduces – and is proud to republish (1973, chapt. DH) – the map of an ideal school (1899, 95); significantly, the library is at the center of the school, close to daily processes, where cognitive needs emerge.

Figure 3: Map of a school (Dewey 1973, 95)

Ranganathan adopted Dewey’s “individual education” methodologies in the classroom, with very good results (Bianchini 2006, 14), but he was essentially misunderstood by his superiors, who judged the endless lists of books he required for his students to be “extravagant”.

Ranganathan’s sudden change of profession, abandoning teaching to become a librarian, can also, to some extent, be attributed to the context of a movement to reestablish schools and libraries, based on “individual education” and “autonomous learning”.

In this way, we can imagine the spirit with which he left India. The principal biographers of Ranganathan highlight the intercontinental journeys (1923-25 and 1948-50) as turning points in his life, presenting many common traits, masterfully synthesized in the conspectus of the book *Philosophy of Library Classification* (Ranganathan 1951b):
• two radical changes in his vocational life (from mathematician to librarian and arrival at retirement);
• two long sequences of visits to innovative projects such as community reading centers and business documentation centers;
• two great crises due to awareness of the inadequacy of the available classification theory for the demands of innovative experiences and of post-war economies;
• two great visions allowing him to overcome these crises, revolutionizing his classification theory: Colon Classification and APUPA pattern;
• two or three decades of patient work after each journey to turn such insights into true classification theories;

The two World Wars changed information flows:

From several sides we tried to dealt with information flows increasing for quantities, specialization and diversity that followed the acceleration triggered by the World Wars and by a growing post-war “economy of team-work on a large scale” (Ranganathan 1951, sect. 286).

The book MicroThought and his Service (Ranganathan 1957) reports difficulties that Frits Donker Duyvis (president of the International Federation of Documentation) encountered in proposing the new classification logic to those who could implement them in new electronic systems.

Duyvis asked for Ranganathan’s help in the search for solutions to the problem of micro-document information overload.

In particular “micro subjects embodied in articles in periodicals or scientific reports” (Ranganathan 1967a, sect. XG14).

Ranganathan’s experience in the United States in 1950 seems equally fundamental. Here too he was invited by UNESCO and the United Nations, as well as by a very large private company, The Rockefeller Foundation. With precise objectives Ranganathan specifies explaining why he accepted the invitation:

It was accepted chiefly to observe documentation in action at close quarters in several industries in the USA. But it was learnt that the purpose of the Foundation inviting me was to explore the role of classification in communication (Ranganathan 1963, 18).

From these elements we can try to make another significant leap in our understanding the APUPA pattern: we now attempt to follow a logical-philosophical plan, adopting an intuitive approach, based on two lexical hints.

3. Penumbra in Philosophy and Psychology: Whitehead, James, Wundt, Leibniz

The first hint is the choice in the APUPA acronym of the Latin calque "umbra" instead of the primary form "shadow". As a classification expert, he was surely aware of the connotation of such a word, which was preserved in the context of specific scientific disciplines.
Two decades earlier, Alfred North Whitehead had built on the term “penumbra” a system of metaphors overturning the representation of consciousness, putting in the center what was usually present only "in shadow";

there is a small focal region of clear illumination, and a large penumbral region of experience which tells of intense experience in dim apprehension….consciousness is the crown of experience, only occasionally attained, not its necessary base (Whitehead 1929, 267).

Thus, he re-issued the challenge of the pragmatist renewal, which four decades earlier William James had issued to the "empiricist" creed of the motherland:

the ridiculous theory of Hume and Berkeley that we can have no images but of perfectly definite things…whilst simple objective qualities are revealed to our knowledge in subjective feelings, relations are not.

On page 258 of "Principles of Psychology" James had written:

If we then consider the cognitive function of different states of mind, we may feel assured that the difference between those that are mere acquaintance…is reducible almost entirely to the absence or presence of psychic fringes or overtones. Knowledge about a thing is knowledge of its relations. Of most of its relations we are only aware in the penumbral nascent way of a fringe of unarticulated affinities about it. The traditional psychology, talks like one who should say a river consists of nothing but pailsful, spoonsful, quartpotsful, barrelsful, and other moulded forms of water. Even were the pails and the pots all actually standing in the stream, still between them the free water would continue to flow. The significance, the value, of the image is all in this halo or penumbra that surrounds and escorts it. Represented by a curve, the neurosis underlying consciousness must at any moment be like this:

Figure 4: The height of the curve above the line stands for the intensity of the brain processes

4. Penumbra in Optics and Astronomy: Kepler, Wundt, Helmholtz

The second hint is based on a counterintuitive pattern: the choice of a progression "from penumbra to umbra", unusual when compared to progression “from shadow to light.” Ranganathan himself often uses "enlightening analogies.” Pragmatist representation also contemplates progression "from umbra to penumbra", but the fruit of the “concrescence process” is the production of something new (named “actual entity”), rather than the recognition of something already existing and its fruits, i.e. “clarification”, “enlightening”.

Consciousness is only the last and greatest of such elements by which the selective character of the individual obscures the external totality from which it originates and which it embodies.
Hence, such counter-intuitiveness reduces; that is, the selective nature of the individual aim obscures the external totality. This is comparable to what happens with “exclude-filters”, where information is created by exclusion from the whole, by obscuring its possibilities. As in the “subtractive color synthesis” process that was studied at that time to understand how the retina works, white light is given off when a surface (or pigment in a painting) reflects all frequencies, but there is total blackness when no frequencies are reflected, while different colors are produced when surfaces retain different parts of the frequency spectrum. As we will see in the next paragraph, the academic discipline of optics was contemporaneously studying these subjects.

The structure of the prehension-apprehension binomial seems to be a clear reference to the Leibnitian “perception-apperception” binomial, reactivated by the progress that the theories of perception had made during those years in Germany; in the 1860s Helmholtz, Young and Max Schultze introduced the "duplicity theory", the discovery that in the retina there were two basic types of photoreceptors, with separate visual functions:

- cones (photopic vision): many small containers, acuity and color vision in bright conditions; one or a few photoreceptors synapse on one neuron;
- rods (scotopic vision): a few larger containers, less colors; sensitivity in dim light, many photoreceptors synapse on one neuron.

Figure 5: “Bell” distribution of receptors on retina

This pattern recalls that which APUPA uses for a common strategy of message selection. The birth of the discipline “psychology” originates from this context. William James had studied Helmholtz, as had Wilhelm Wundt, who in Leipzig managed the first laboratory devoted to the study of mental processes and experimentally measuring them. In the city of Leipzig, two centuries earlier, Leibniz had anticipated this approach with the concept of “small perception”.

Yet, the continuum influencing development of the penumbra idea seems to go even further back.

In 1604, Kepler had written an important book about the application of optics to astronomy entitled “Ad Vitellionem paralipomena quibus astronomiae pars optica traditur.” The possibility of measuring the trajectories of celestial bodies is strongly linked with the ability to control perceptive distortions.

A penumbra causes a distortion: bodies appear larger than they really are, leading to inaccurate measurements.

Figure 6: Penumbra effect: bodies appear larger

It does not seem a coincidence, but rather another hint, that the English term “umbra” has a secondary meaning, whose presence is also recorded from the beginning of the 1600s: the image of "an uninvited guest accompanying an invited one" – an uninvited larger silhouette, as Kepler might say.

To express the intermediate degree of the shadow on the moon during eclipses and to describe this sort of “relational interference” between macro (celestial giants) and micro (human ocular bulbs), Kepler had coined a new term – *penumbra* – juxtaposing the Latin terms pæne (quasi) and umbra (shadow).

The perceiver’s form and movements can be identified through relations with other bodies in the form of a reported shadow. In lunar eclipses we gather information on the obscuring form of the Earth, whose silhouette becomes visible in so far as it blocks out the blinding sunlight, which, no longer reaching the lunar surface, does not bounce back to the retinal back wall of our "ocular cave." Moreover, the shadow is indeed a sign not of emptiness, but of the fullness that determines it, and of its relations with other bodies in which we can see ourselves reflected.

And here we return to the importance that these philosophical considerations could represent for those who have tried to tackle the problem of overload.
5. Logics and Mathematics: umbral and relational calculus


Miksa’s (1997) periodization is founded on the role played by mathematics in Ranganathan’s classification theory and particularly the introduction of mathematical tools in classification schemes. He emphasizes how the second period disclosed the flaws of predetermining a rigid facet formula. A new analytical-synthetic model is taken up, with all the potential allowed by the latest mathematical innovations and in automatic and flexible sorting.

Ranganathan tries to connect the philosophical with the mathematical level contained in Russell-Whitehead’s research.

In a certain sense everything is everywhere at all times. For every location involves an aspect of itself in every other location. Thus every spatio-temporal standpoint mirrors the world (Whitehead 1945, 114).

This led Ranganathan to the idea of a “universe of subjects” articulated in infinite dimensions. Ranganathan and the Library Research Circle assiduously studied Whitehead-Russell’s non-metrical projective geometry, in an attempt to bring about a linear mapping of multidimensional knowledge (La Barre 2006), by continuing Otlet’s challenge:

Ranganathan does follow this line of reasoning of successive syntheses in his loops of “micro-thought.” However, whereas Otlet thought that this development could be handled by an update of the UDC from time to time, Ranganathan was convinced that it implied a fundamental change in classification (van den Heuvel 2011, 110).

Ranganathan (1951b, sect. 47) declares:

The representation of any continuum on a continuum of a smaller number of dimensions will admit of alternatives. Take for example the simple case of representing three dimensional space on two dimensional space—the representation of a globe as a plane map.

The mathematical concept of “transformation” recalls the use of the term “transformation” to describe the passages from one phase to another in the model of communication processes represented in Ranganathan (1951a sect. 3212).
In the last quarter of the 19th century, John Hopkins University in Baltimore became the mathematical center of the USA. In 1877 the world-renowned authority James Joseph Sylvester began coordinating the mathematics community. From 1879 until 1884, Charles Sanders Peirce held the position of lecturer on logic, on a recommendation from his great friend William James. In the first half of 1882, the university hosted Arthur Cayley (Murphey 1993). The authority of the professors attracted brilliant students such as John Dewey.

Peirce, in a 1903 article, refers to one of his usual disagreements with Sylvester: a controversy on umbral calculus, a type of mathematical notation that allows description of the properties of a universe of values contained in a matrix through a function that Sylvester first called “umbra”.

Ranganathan uses the concept “universe of subjects” to define the growing number of isolates that the classification formula can select from, combining the values of the multiple facets of a book.

Peirce first criticizes the umbral metaphor:

Sylvester's name umbra...must, I fear, be retained, although ion or radicle would be far better. For who ever heard of two shadows combining together to form a substance! In other mathematics, they have no existence in the universe of quantity. But joined together in sets they do. They are just like chemicals radicals, each having a certain number of unsatisfied wants. When each of these is satisfied by union with another, the completely saturated whole has an existence in the universe of quantity.

Peirce affirms that Sylvester was not the first to use these techniques:

What Sylvester called ‘my umbral notation’ had first been published in 1693 by another man of some talent, named Godfry William Leibniz (Peirce 1997, 125).

An explication of the basic concept of relational calculus (Smith 2012) highlights its
relationship with the theory of “small perception”. Let’s consider two right triangles that have an apex, Z, in common. Since the two triangles are similar, it follows that the ratio \( y/x \) is equal to \((Y-y)/X\) and it will remain equal to it. Also, if the straight line EI always preserves the same angle Z, it increasingly approaches point F. What happens when the straight line EI passes through F itself? The point Z and E will fall directly on F, then the straight lines x and y will vanish; they will become equal to zero. But they still maintain an algebraic relation to each other, which is expressed in the relation of X to Y.

Figure 8: Leibniz’s divulgative graphic to explain relational calculus

At the end of the 19th century, Gottlob Frege had reached Leibniz’s objective of *characteristica universalis*, a new language based on *ars combinatoria*. According to Menon (2008), Ranganathan assumed the challenge:

la détermination des isolats se rapproche d’une recherche des concepts “séminaux”, ou notions primitives, telle qu’elle était envisagée dans le cadre de la caractéristique universelle.

Srinivasa Ramanujan, one of the greatest exponents of umbral calculus studied in Madras with Ranganathan’s teacher and mentor E. B. Ross (Ranganathan 1967a, DB06). When he died he was working to “mock modular form”, formulas characterized by an associated modular function ("its shadow"). The search for those precious leaves was among the reasons Ranganathan had traveled to England in 1924.

I’m looking for mathematicians for possible future research.

**Conclusion**

Throughout this research I appreciated the surprising relevance of the analysis that led to the foundations of KO for actual societal challenges; speed, intensity and complexity of workflows experienced during and after the Second World War seems now to have extended to society as a whole.
Ranganathan’s dream when faced with a “Meccano” box a century ago is a piece of our near future: Let’s prepare for the centenary.

Acknowledgements
This research arose from a proposal from Claudio Gnoli during a seminary held by Carlo Bianchini (Bianchini, C, Giusti, L., Gnoli, C. 2017). The author thanks C.Albin, F.Giusti, V.Lo Faro, A.Merletti, B.Muraca, F.V.Sbrana, L.Testoni, N.Vallinoto.

References
Bianchini, Carlo, Giusti, Luca, Gnoli, Claudio (2017) The APUPA bell curve: Ranganathan’s visual pattern for knowledge organization”. Available online at Les Cahiers du numerique, 13(1).


Eva Hourihan Jansen

Photography as a legitimate technique for domain analysis in Knowledge Organization

Abstract
In KO research, it is broadly understood that visualization offers the advantage of providing a graphic overview of a domain (Smiraglia 2014, 95). This paper argues that in addition to citation analysis and other methods of visualizing domains, photography is a legitimate data collection and analysis technique for accomplishing KO research. Findings from a study of occupational classification, in the context of employment support for newcomer professionals, demonstrate that photography can help to generate themes and describe the character and texture of a domain as well as what influences it.

1. Introduction
Recently, the Sixth North American Symposium on Knowledge Organization (NASKO 2017) contemplated visualizing knowledge, knowledge organization, and knowledge organization systems. Clearly, different forms of visualization can lead to navigational maps and some recent examples include citation analysis (Smiraglia 2017 in ibid.), cladistic visualizations (Campbell and Mayhew, 2017 in ibid.), knowledge graphs (Zhao, Ma and Xia 2017 in ibid.) meta-theoretical visualizations (Araujo, Tennis and Guimarães 2017 in ibid.), along with node-link diagrams and cover images (Hook and Gantchev 2017 in ibid.). As a collectivist approach that recognizes experience is shaped by social and cultural communication, domain analysis takes the unit of analysis beyond the individual to the group level and looks toward characteristics of the environment (Hartel 2003). Hence, in knowledge organization (KO) research, among the techniques for visualizing domains that appear most often in the literature, citation analysis is considered a valid form of visualization in KO (Smiraglia 2014). Visualization offers the advantage of providing a graphic overview of a domain (Smiraglia 2014, 95). This paper describes data collection and analysis techniques to position photography and photographs as another useful technique for accomplishing KO research.

2. Context
The context of this investigation is a completed study of Canada’s national occupational classification (NOC), part of a genre of state sponsored standard classification systems used in labour markets around the world. The study sought to describe norms, values and expectations expressed in relation to the NOC, and asks how these may be represented to people interested in knowledge organization systems and information studies. The study was set in a not-for-profit organization that supports the provision of services to newcomer professionals within a network of community
service providers and employers in the local labour market. As a KO system that is constructed through consultation with a broad group (Howarth and Hourihan 2014), it therefore demands a collectivist approach like domain analysis.

3. Methods

The study is characterized by fieldwork methods (Sandstrom and Sandstrom 1995) that comprised documenting participant observation (Baker 2005; Spradley 1980); through ethnographic interviews (Spradley 1979) and photography (Hartel and Thomson 2011). Participant observation involved the researcher in regular attendance at the fieldwork site over a four-month period to learn about the ways occupational classification is operationalized amidst organizational activities. During this time the researcher gathered over 150 photographs.

The researcher took photos largely spontaneously throughout the field as long as discretion and maintaining trust permitted. Each day of fieldwork was consistently and discretely documented in photographs taken with a mobile phone. Each evening after fieldwork photos were downloaded on the researcher’s personal computer. The corpus of photographs was also professionally printed on photographic paper. The intent was to visually document her own presence in the research setting as a way to enrich fieldnote writing practices (Emerson, Fretz and Shaw 2011) and to expand the dataset by eliciting something akin to alternative documents (Weissenberger, 2015). Through analysis of images and text the role of photography as a technique for eliciting themes from the body of research data became increasingly relevant.

4. Photographs in data collection & analysis

Findings pertinent to contemplating the researcher’s role as a participant observer in this case were revealed through photos taken amid constraints of privacy and non-disclosure. Although the photos were collected with sensitivity to tips recommended by Hartel and Thomson (2011), they were at times secretive or seemingly non-sensical.

Later, when participant observation was complete and the researcher had removed herself from the field, the rich detail and diversity of the photographic inventory led to inventing several codes for the data. Sorting the corpus of 150 photos was integral to developing analytical themes throughout the writing process. Photos were coded in conjunction with the other written and collected documentation that formed the data set.
Among emergent themes that I identified, were those that relate well to Lee’s (2003) framework of immediate, adjacent and outside space, where spaces are conceptualized as locations where “a great quantity of material information sources” is contained (ibid. p. 427). This framework became analytically useful as I sorted through the photographic and textual documentation produced in this study. Preliminary analysis of photos aided in identifying themes, in particular locating the NOC among a variety of personal, organizational, and institutional spaces.

The least obtrusive and most abundant photos in the inventory feature personal workspace. These spaces, exemplars of which are featured in Figures 2 and 3, are characterized by documents, desks, screens indicate common structures of office work.
Thematically, the images above were coded initially as *spaces inside the office*. As an ethnographer documenting fieldwork, photography proved to be a helpful technique for documenting *immediate information space* (Lee 2003; Hartel and Thomson 2011) and reflecting on the nature of this role and its relationship to the research data. The photos demarcate the space of participant observation. It is where up-close analysis of the NOC took place, among documents, reports, spreadsheets, on the local network.

This up-close work was eventually differentiated, by means of visual analysis, from other types of fieldwork encounters. Images of meeting spaces, and hallways indicate additional structures within where I participated. Through inductive analysis I grouped these photos with fieldnotes depicting meetings, conversations and interviews. Figures 4 and 5 feature meeting spaces away from the workstation but contained in the organization’s offices.

![Figure 4: Meeting room](image1.png)  ![Figure 5: Kitchen](image2.png)

Here photos deepen recall of the sociality of the office space where, for example, colleagues shared travel stories in the kitchen over snacks. Standing together at the map of local restaurants, we read short-hand reviews and recommendations for local lunch places. Together with employment coaches from partner agencies I learned about using the “IT system” and “coding” mentees with the NOC. These photos feature typical materials and spaces of office workplaces ranging from tables and chairs to microwave ovens, laptops, notepads, maps and sticky notes. Unlike the workstation photos, they
depict situations where people gathered together. Through a combination of inductive and deductive analysis these spaces, I began to relate these shared areas to the notion of adjacent spaces (Lee 2003).

Photographs that were more difficult to capture include those I took at offsite events. I began by collecting photos of rooms where events took place. These I collated with jottings taken at events and fieldnotes written up after events. In one instance attendance at a mentoring event was lower than expected because newcomers were unable to find the location. This signaled a need to modify my approach. I began to take photos of getting into the fieldsite, in order to reflect on the things I take for granted about navigating the local labour market. I had taken for granted that I know how to find places and how much time under normal circumstances to allow for in order to reach a destination. This data collection effort produced some mundane pictures of public transit escalators, passageways, lobby areas and street corners as seen in Figures 6 and 7. Figure 6 features a decorated Christmas tree and an iconic winter scene in stained glass. Figure 7 features scene I came upon at a busy intersection in the city.

As the weeks went on I modified my approach and began to take photos of my journey to fieldwork. Events took place in space donated to the employment council by corporate partners. Some meeting spaces were in bank buildings in the city’s downtown core. Other times events were held in space at a local university or a public television studio. My initial code for photos taken on the way to fieldwork was transitional space. While some of these spaces were adjacent to fieldwork events, and literally outdoors,
it remains unclear to me if or how they fit the notions of adjacent or outside information spaces.

Finally, my initial analysis lead to grouping photos of events attended by but not organized by members of the fieldsite organization. For example, hosting a table at an employment information fair for newcomers to Canada and a discussion panel concerning changes to provincial immigration policy are featured in Figures 8 and 9.

Figure 6: Employment Fair

Figure 7: Community Centre Panel

Photographs served as an analytical resource for elaborating on the characters and textures of these spaces. In a sense, Lee’s concept outside information space may help delineate comparisons with the happenings taking place among inside, adjacent, and transitional spaces that inform this domain. Photos assembled under this theme feature chairs, tables, audiovisual media and food catering, the typical accoutrements of corporate workspaces in a large city. When analyzed in relation to my fieldnotes, I began to draw distinctions among private and public spaces, to ponder the dimensions of accessibility of such work-related spaces. Were it not for my role as participant observer, it would be unlikely experience for me to see the interiors of these spaces or to interact with the people there.

4. Discussion

In knowledge organization, research citation analysis and concept mapping are techniques that offer insight into the social context of research. Part of their appeal aligns with objectivity, and as such they effectively represent what Talja et al. (1999) suggest is a normative approach to naming entities at play in a domain. Interpretive approaches to photographic data can highlight the complexity of studying knowledge
organization. This collection of photographs, organized along different dimensions becomes a resource for not only taking inventory of materials present in spaces, but also for depicting social situations and events in the research process and experience. To that end it expands the experience of the researcher toward different possible viewpoints. Simple comparisons abound, including how private desktop space may be contrasted with the open space of a social event. The private office space may be compared to the private meeting space. The sociality of a webinar may be compared to the sociality of a mentoring event. The concrete architecture featured across the photos may be synthesized with the abstract architecture of events and encounters. Photography enriches the conceptual grounds for interpreting the knowledge organization system, the object of study. In effect, adding a corpus of photographs to the dataset aided in pulling the researcher’s focus outward from term-based and activity-based analysis toward the more challenging task of social analysis. The important up-close view of classification work is made relevant to socio-cultural analysis by adjusting the aperture of the view away from the text and terms, away from the cognitive focus of activity to focus in on broader spaces of action.

5. Conclusions

Participant observation is a method enriched by photography. Photographs simultaneously act as documents and memory aids in fieldwork; act as analytical resources for evoking themes and producing a narrative description of knowledge organization practices; and play a role in communicating research findings.

Photography is an important technique featured in studies of immediate information space (Hartel 2007; Thomson 2010). Andrade, Urquhart and Arthanari (2015) claim researchers could make better use of images as a source of data by increasing the use of visual material, by systematically unpacking visual data and by collecting both primary and secondary visual data for analysis as well as later presentation. In their highly influential monograph, Bowker and Star advocated broadening the theoretical fabric of studying categories to studying large-scale systems where the analysis of classifications “knits together to form the texture of a social space” (1999, 286). Fully cognizant of the challenge of seeing relationships among people, things, moral order, categories and standards, they suggest a good map and compass are needed. The map and compass offer an apt metaphor for research design and theoretical framework. To this metaphorical tool kit, I propose that we had a camera. The camera, in the metaphorical sense, affords the methodological inversion needed to study extant knowledge organization systems. Whereas the range of visualizations that exist to produce maps of scholarship have come to define visualization in KO research, I propose that a camera, taken into the information spaces of the researcher can re-invigorate the research experience by offering additional means for visualizing research data. This study also suggests photography is a legitimate technique for domain analysis.
and can contribute to extending knowledge organization research through methodological diversity.

This research was supported by the Ethel Auster Scholarship & Bursary Award and the University of Toronto.

References


Michael Kleineberg

Reconstructionism: a comparative method for viewpoint analysis and indexing using the example of Kohlberg’s moral stages

Abstract
This contribution introduces Jürgen Habermas’s methodology of hermeneutic reconstructionism to the field of knowledge organization in order to provide a comparative tool for viewpoint analysis and indexing. Hermeneutic reconstructionism is a validity-seeking form of interpretation that is based on a rational reconstruction of the implicit know-how of language users in terms of cognitive presuppositions and communicative competence. Using the example of Lawrence Kohlberg’s reconstruction of developmental stages of moral consciousness, this paper seeks to indicate how documents that express different moral views can be analyzed and indexed according to the author’s form of moral judgment or justice reasoning.

1. Introduction: from subject indexing to viewpoint indexing
The widely accepted concept of epistemic pluralism calls for a “multi-perspective knowledge organization” (Kaipainen and Hautamäki 2011, 509) that is able to identify and interrelate different viewpoints on the same subject matter. Therefore, indexing theory is shifting the focus from traditional subject analysis, which is concerned with the aboutness of the more or less explicit semantic content of a document, to a kind of viewpoint analysis that seeks to reconstruct the often only implicit perspective taken by the author (Crowe 1986; Andersen and Christensen 2001; Biagetti 2006; Kleineberg 2013). In semiotic terms, viewpoint analysis and indexing refers to the “pragmatic aspect” (Hutchins 1975, 4) of language or sign systems as a relation between a symbolic expression and its users, that is, the context of meaning production. Following Ludwig Wittgenstein’s linguistic pragmatics, Andersen and Christensen (2001) emphasize that knowing the document is not sufficient to understand its meaning. Instead, an analysis of the underlying language game is required, which is directed to an implicit know-how rather than an explicit know-that. This is the point of departure for Jürgen Habermas’s (1990, 28) “hermeneutic reconstructionism” that is based on a rational reconstruction of the author’s intuitive rule knowledge of language games or communicative actions. In contrast to a strong contextualism resulting in a relativist “hermeneutic nihilism” (Campbell 1988, 505), Habermas’s approach offers a comparative method for analyzing mutually contradicting perspectives or ways of reasoning within an inclusive framework (Cooke 1994; Bookman 2002; Backlund 2005; Pedersen 2008). This paper presents the example of Lawrence Kohlberg’s rational reconstruction of moral development and its application by hermeneutic-reconstructive viewpoint analyses in historical and cross-cultural studies on ethics and morals (Radding 1978; Schluchter 1981; Apel 1988; Rosenberg, Ward and Chilton 1988; Roetz 1993; Hallpike 2017).
2. Hermeneutics and rational reconstruction

Indexing theory is concerned with the meaning of a document. This involves hermeneutics as the theory or methodology of interpretation and understanding of symbolic expressions. In opposition to what Habermas (1990, 29) calls a “radical hermeneutics” abandoning the claim to explanatory knowledge, his hermeneutic reconstructionism presents a kind of “validity-seeking hermeneutics” (Campbell 1988, 505) that maintains the distinction between socially agreed upon and valid points of view. Although knowledge is always culturally and historically embedded, validity claims of truth, rightness, and truthfulness transcend all given contexts since they rely on disputable reasons that have the subversive potential to call into question all spatio-temporally defined social agreements (Bookman 2002; Maeve 1994). According to Habermas (1990, 30 [emphases original]), any interpretation is to some extent a rational interpretation in which reasons are evaluated according to presumably universal standards of rationality:

But only to the extent to which the interpreter also grasps the reasons why the author’s utterances seemed rational to the author himself does he understand what the author meant. The interpreter, then, understands the meaning of a text only insofar as he understands why the author felt justified in putting forth certain propositions as being true, in recognizing certain values and norms as being right, and in expressing certain experiences (or attributing them to others) as being authentic.

Habermas (1979) distinguishes two modes of meaning explication. The first mode investigates the surface structure, that is, the semantic content of a symbolic expression or the author’s explicit know-that. In indexing theory, this traditional form of interpretation is concerned with subject analysis and the aboutness of a document. The second mode, however, is related to the pragmatic context in which a document is produced. It directs itself to the deep structure, that is, the generative structures according to which a symbolic expression has been brought forth, meaning the author’s implicit know-how that involves cognitive presuppositions and communicative competence.

With regard to such a “depth hermeneutics” (De Mul 1997, 240), Habermas (1979) makes a methodological distinction between empirical pragmatics (e.g., sociolinguistics), which investigates context-bound utterances, and universal or formal pragmatics, which seeks to reconstruct general patterns of communicative action that apply to any language in any context. In Habermas’s formal-pragmatic approach, communicative competence, defined as the ability to employ sentences in speech acts, is considered to be the result of a learning process that follows a logic of development in that it runs through an invariant sequence of hierarchically ordered generative structures, regardless of the context-dependent mechanisms or dynamics of development. Furthermore, the explanation of such a learning process is able to avoid the naturalistic fallacy, which states that an “ought” cannot be deduced from an “is,” if normative-philosophical rational reconstructions are complemented to empirical-
psychological analyses, as can be found in the work of philosopher-psychologists like John Dewey, James M. Baldwin, George H. Mead, Jean Piaget, or Lawrence Kohlberg.

3. Viewpoint analysis and indexing as exemplified by Kohlberg’s moral stages

The process of traditional subject indexing is based on two main steps. The first step is an analysis of the semantic content or the aboutness of a document using the terminology of the author and the second step is a translation of the analyzed subject matter into index terms of a controlled vocabulary. Viewpoint indexing, however, requires a different mode of meaning explication since the author’s underlying generative structure is usually not made explicit in the document. The advantage of hermeneutic reconstructionism is its recourse to already existing rational reconstructions of domain-specific competences (e.g., mathematico-logical reasoning, physical thinking, spatial representation, aesthetic judgment, ego identity, self-understanding, social cognition, moral consciousness, religious thought) that can be regarded as specifications of the general developmental-logical organizing principle of integrative levels of knowing (Kleineberg 2014; 2017).

One of the best-known examples is presented by Kohlberg’s moral stages derived from a “rational reconstruction of the ontogenesis of justice reasoning” (Kohlberg, Levine and Hewer 1983, 10). From a formal-pragmatic perspective, moral consciousness dealing with values and norms refers to regulative speech acts in relation to the social world claiming validity in terms of rightness (Habermas 1990). In this respect, moral consciousness is a universal and cross-cultural phenomenon. From a developmental-logical perspective, however, stage-like qualitative differences become apparent in the justification of what counts as morally just and reasonable (see Table 1).

Kohlberg and colleagues (1983) distinguish functional stages that are related to experiences of new socio-cultural roles of the maturing person (e.g., Erik Erikson’s identity development), soft structural stages that focus on the form of development and involve a consciously reflecting ego or self as source of a totalistic meaning-making or worldview (e.g., Jane Loevinger’s ego development), and hard structural stages that are limited to discrete domains of such worldviews (e.g., Piaget’s cognitive development). The methodological benefit of hard structural stages, like Kohlberg’s moral stages, is the precision in articulating the logic of development that exhibits the following formal properties (Kohlberg, Levine and Hewer 1983, 31 [emphasis original]; see also Kleineberg 2014; 2017; Lourenço 2016):

1. Stages imply a distinction or qualitative difference in structures (modes of thinking) that still serve the same basic function (for example, intelligence) at various points in development.
2. These different structures form an invariant sequence, order, or succession in individual development. While cultural factors may speed up, slow down, or stop
(3) Each of these different and sequential modes of thought forms a ‘structural whole’. [...] The implication is that various aspects of stage structures should appear as a consistent cluster of responses in development.

(4) Stages are hierarchical integrations. As noted, stages form an order of increasingly differentiated and integrated structures to fulfill a common function. Accordingly, higher stages displace (or, rather, integrate) the structures found at lower stages.

Table 1: Kohlberg’s developmental stages of moral consciousness

<table>
<thead>
<tr>
<th>Moral stage</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level A</strong></td>
<td></td>
</tr>
<tr>
<td>Preconventional</td>
<td>Action and judgment are directed by a naive hedonism and not by internalized expectations of others, group solidarity, or general ideas of justice. Sanction: punishment (deprivation of physical rewards).</td>
</tr>
<tr>
<td>Stage 1</td>
<td>Punishment and obedience orientation The physical consequences decide whether an action is right or wrong. Obedience to authority is guided by the egocentric interest in benefits and the avoidance of physical harm. Egocentric perspective: complementarity of order and obedience.</td>
</tr>
<tr>
<td>Stage 2</td>
<td>Instrumental exchange What is right, is, like at Stage 1, that which satisfies one’s own immediate needs while strategically also recognizing the interests of others. The orientation is towards elementary rules of reciprocity and concrete exchange like “you scratch my back and I’ll scratch yours.” Egocentric perspective: symmetry of compensation.</td>
</tr>
<tr>
<td><strong>Level B</strong></td>
<td></td>
</tr>
<tr>
<td>Conventional</td>
<td>Action and judgment are directed in an alienated way (especially at Stage 4) by means of self-identification with heteronomous rules and norms. The aim is not only adjustment to a given order, but its maintenance for its own sake. Sanction: shame (withdrawal of love and social recognition).</td>
</tr>
<tr>
<td>Stage 3</td>
<td>“Good boy—nice girl” orientation The right is what “pleases” and what is motivated by the rules, values, and expectations of groups. Primary group perspective: conformity to roles.</td>
</tr>
<tr>
<td>Stage 4</td>
<td>Law and order orientation The right is what is in accordance with the prescriptions of one’s role in state and society. Action and judgment are motivated by the respect for authority and doing one’s duty as a value per se. Perspective of a collectivity (the system’s point of view): conformity to the existing system of norms.</td>
</tr>
<tr>
<td><strong>Level C</strong></td>
<td></td>
</tr>
<tr>
<td>Postconventional</td>
<td>Conventional morality is not necessarily rejected (except at State 4½), but its acceptance is accompanied by a clear awareness of the possible conflict between conventions and law on the one hand and morality on the other. Sanction: guilt (reaction of conscience).</td>
</tr>
<tr>
<td>Stage 4½</td>
<td>“Anything goes” and youthful protest What is right is a question of arbitrary subjective decisions. This stage is characterized by a radical rejection of the alienated conventionalism of Level B and the recourse to the naive pleasure principle of Level A. Instead of new normative rules, this stage proclaims a provocative “beyond good and evil.” It is postconventional but not yet principled.</td>
</tr>
</tbody>
</table>
Stage 5
Utilitarian, relativistic
social contract
orientation

What is right is not predetermined, like on Level B, by the existing institutions and social conditions. It is first of all a matter of personal, relative values and opinions. Beyond this it is defined in terms of standards that have been agreed upon by free and equal individuals and that can be changed by regulated procedures. Principled perspective (prior to society): orientation toward principles of justice.

Stage 6
Universal ethical
principle orientation

The right is what is in accordance with abstract, consistent, and universally valid principles. It is based on the autonomous decision of conscience. Procedural perspective (ideal role taking): orientation toward procedures for justifying norms.

Source: adapted from Roetz (1993, 26-27) and Habermas (1979, 81; 1990, 167)

There are several modified formulations of Kohlberg’s moral stages (Kohlberg, Levine and Hewer 1983; Reed 1997; Gibbs et al. 2007). For example, Stage 4½ was initially regarded as a transitional stage but its stabilization can be found in philosophical versions of value skepticism from the axial age, exemplified by Chinese Daoists or Greek sophists (Roetz 1993), to postmodernism inspired by Nietzschean nihilism (Habermas 1990; Apel 1988). From an empirical point of view, Stage 5 reasoning seems not to be developed in all cultures or in the majority, for instance, of the US adult population (Gibbs et al. 2007), while Stage 6 reasoning is not detected at all by longitudinal studies and seems to be represented only by a small elite sample including thinkers like Mohandas Gandhi and Martin Luther King, or theorists of universal ethics of speech (Kohlberg, Levine and Hewer 1983). Occasionally, a metaphorical Stage 7 is considered in terms of a cosmic or infinite perspective attributed, for example, to Marcus Aurelius or Baruch Spinoza, but such a quasi-religious soft structural stage cannot claim universality since a rational reconstruction presumes a developmental endpoint upon which rational agents could agree (Kohlberg and Ryncarz 1990; for alternative reconstructive methods see Wilber 1999; Dux 2004).

Despite some controversial issues, Gibbs and colleagues (2007, 491) conclude in their meta-analysis that “Kohlberg was in principle correct regarding the universality of basic moral judgment development.” The decisive point is that even though there is a plurality of perspectives on morality or justice, the structural development of moral judgment seems to follow an invariant sequence of stages across cultures and sex differences (Kohlberg, Levine and Hewer 1983; Jorgensen 2006; Gibbs et al. 2007). This offers a non-relativistic comparative framework for a systematic organization of different moral views, as underlined by Habermas (1990, 117 [emphases original]):

Opponents of universalistic ethics generally bring up the fact that different cultures have different conceptions of morality. To oppose relativistic objections of this kind, Kohlberg’s theory of moral development offers the possibility of (a) reducing the empirical diversity of existing moral views to variation in the contents, in contrast to the universal forms, of moral judgment and (b) explaining the remaining structural differences between moralities as differences in the stage of development of the capacity for moral judgment.
Without drawing too hasty parallels, one might expect some “homologous structures of consciousness in the histories of the individual and the species” (Habermas 1979, 99; see also Wilber 1999; Barnes 2000; Bammé 2011; Bellah 2011; Dinzelbacher 2015). Accordingly, viewpoint analyses based on Kohlberg’s moral stages are also provided for collectively shared interpretive systems like religious and philosophical ethics (Roetz 1993; Apel 1988; Hallpike 2017) or institutionalized legal and moral representations (Radding 1978; Schluchter 1981; Rosenberg, Ward and Chilton 1988; Oesterdiekhoff 2014). However, it cannot be emphasized enough that hermeneutic reconstructionism is not about judging individuals or cultures as a whole; instead, it is about an evaluation of particular validity claims raised in communicative actions, including by means of written documents (Backlund 2005; Budd 2011; see Table 2).

Table 2: Examples of viewpoint analysis and indexing using Kohlberg’s moral stages

<table>
<thead>
<tr>
<th>Document</th>
<th>Significant features</th>
<th>Moral stage</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shujing (Book of Documents), chapter Pan Geng, c. 1100 BCE</td>
<td>Orientation towards role and upholding of the system instead of preference for kin. Establishment of a political state order.</td>
<td>Stage 4</td>
<td>Roetz (1993, 35)</td>
</tr>
<tr>
<td>Iliad, attributed to Homer, 8th century BCE</td>
<td>Thoughts being placed in the mind by gods. Lack of rights for individuals. Heroism and despotism based on power.</td>
<td>Stage 2</td>
<td>Barnes (2000, 107-108)</td>
</tr>
<tr>
<td>Recorded sayings of Buddha, Confucius, the Jewish prophets, and Muhammad, 6th century BCE to 7th century ACE</td>
<td>Reflection on society as a whole, on individuals independently of social status. Distinction between moral principles and duties or customs. Importance of conscience. Significance of intention.</td>
<td>Stage 5</td>
<td>Hallpike (2017, 301)</td>
</tr>
<tr>
<td>Xunzi, 3rd century BCE</td>
<td>The state as product of reasonable deliberation. Utilitarian rationale. Critique of conventional role-morality: “Follow the Dao and not the ruler, follow justice and not the father.”</td>
<td>Stage 5</td>
<td>Roetz (1993, 63-65, 269, 274)</td>
</tr>
<tr>
<td>The Social Contract by</td>
<td>Critique of feudal system. Inalienable rights</td>
<td>Stage 5</td>
<td>Bammé</td>
</tr>
</tbody>
</table>
Jean-Jacques Rousseau, 1762

United States Declaration of Independence, 1776

Critique of Practical Reason by Immanuel Kant, 1788

Beyond Good and Evil by Friedrich Nietzsche, 1886

A Theory of Justice by John Rawls, 1971

of liberty. Principles of democracy based on natural law and social contract.

Inalienable rights based on equality and mutual respect. Adjudicating public law derived from rational debate and consent.

Ethic of responsibility in monological, not yet dialogical form. Validity based on reflexive principles. Evaluation of consequences of conduct.

Denunciation of validity claims of morality. Genealogy as naturalistic reductionism.

Moral principles designed to reach agreement in situations of potential moral conflict. Justice as reversibility.

(2011, 566)

Stage 5

Rosenberg, Ward and Chilton (1988, 152)

Stage 5

Schluchter (1981, 63-64)

Stage 4½

Apel (1988, 388)

Stage 6


4. Conclusion: towards a multi-perspective Knowledge Organization

In opposition to strong contextualist and relativist approaches to the explication of meaning, hermeneutic reconstructionism offers a comparative method for viewpoint analysis and indexing. Different modes of thinking are not simply considered to be equally valid but to be amenable to an evaluation of reasons according to normative standards of rationality. The organizing principle of integrative levels of knowing, exemplified in this paper by Kohlberg’s moral stages, can also be applied to the development of other domain-specific competences and offers a powerful tool for expressive documentary languages with regard to viewpoint representation (Kleineberg 2014; 2017). Moreover, a comprehensive multi-perspective knowledge organization should not only consider the implicit know-how of authors but also of readers, such as indexers and users of information systems, in terms of levels of interpretation and understanding (Kreft 1977; Beak 2014).

References


Abstract
Relations between concepts (and/or entities, events, and other things) vary depending on the criteria by which relations are defined or viewed. In the domain of research data management, different types of research generate different types of data and terminologies vary between practitioners and basic science researchers even within the same disciplinary domain. Interactions between datasets, between datasets and documentation, and between datasets and computing code can result in different types of relations. This paper employs a framework of analysis to study concept relation types in the research data management domain. By using two cases – one is the GenBank annotation records and the other is the data and artifact collection from a gravitational wave search, this paper demonstrates the types of relations existing in and between datasets, publications, computing codes, and workflows. The analysis and generalization of these relations references the research in AI’s knowledge representation and knowledge organization systems (KOS), including both ad hoc subject categories and formal KOS, because in the next AI era, relations as one of the key components of AI applications will be required to function not only as part of KOS for indexing data and publications, but more importantly, to function as codifiable knowledge for machine consumption.

Introduction
Research data management (RDM) is increasingly becoming part of routine work for many libraries and research institutions. Vocabularies for RDM have also emerged to meet the needs of data management and curation. Examples include a taxonomy of data-related terms for the geographical information systems (GIS) (Cole 2005), the vocabulary for research data repository registration and description (Vierkant et al., 2012), and the DPCVocab framework by Chao et al. (2015). While these endeavors provide useful frameworks for understanding the description of data and curation practices, the complex and entangled relations between concepts in datasets and other research artifacts are still a less known territory waiting for exploration. The recent hype in artificial intelligence (AI) research and development and the promises that knowledge organization systems (KOS) have toward AI call for a re-examination of relations in KOS.

This paper is an extension of this author’s 2002 ISKO article on the evolving paradigm of knowledge representation and organization, in which the author presented two opposite but crossover spectra in knowledge representation and organization, one for pragmatism vs. epistemologism and the other for integration vs. disintegration (Qin 2002). By using two cases – one is the GenBank annotation records and the other is the data and artifact collection from a gravitational wave search, this paper will demonstrate the types of relations existing in and between datasets, publications, computing codes, and workflows. The analysis and generalization of these relations
will reference the research in AI’s knowledge representation and KOS, including both *ad hoc* subject categories and formal KOS, because in the next AI era, relations as one of the key components of AI applications will be required to function not only as part of KOS for indexing data and publications, but more importantly, to function as codifiable knowledge for machine consumption.

**Framework of analysis**

Concepts and relations between concepts are the backbones of knowledge organization system (KOS). The most common types of relations among terms in thesauri and subject heading lists are broader terms (BT), narrower terms (NT), and related terms (RT), which establish parent-child, part-whole, or associative relations between concepts. Such relations primarily represent the scope of concepts and can be limited in representing many other concept relations beyond concept scope. For example, the Unified Medical Language System (UMLS) defines about 60 relations in five broad categories: R1. physically_related_to, R2. spatically_related_to, R3. functionally_related_to, R4. temporally_related_to, and R5. conceptually_related_to (UTS 2017). Another example is the relationships defined by Schema.org for Dataset, which are largely inherited from the CreativeWork class. Major relationships for Dataset (or CreativeWork) include hasPart, isBasedOn, isPartOf, and mentions (Schema.org 2017).

The content and formats of a KOS are largely determined by the purpose of KOS and the technology available at the time. Thesauri and subject heading lists fit into early computing technology, the purpose of which was to assist in indexing documents. Term-centric KOS worked well to meet the needs for indexing documents, but in the digital era where information and data are highly interconnected, the term-centric approach becomes too limited to fulfill the increasing demand for effective automatic representation of documents and data.

Ontologies as one of the KOS types entail much richer relations for the concepts they cover. In both the UMLS and Schema.org examples, the relations between concepts have gone far beyond the traditional BT/NT/RT types. Relations between concepts (and/or entities, events, and other things) vary depending on the criteria by which relations are defined or viewed. In the domain of research data management, it is known that different types of research, *e.g.*, experimental, observation, simulation, and survey, generate different types of data and terminologies vary between practitioners and basic science researchers even within the same disciplinary domain. Datasets often come with documentation (some in the form of user guide or manual) and computing code, and are associated with publications. Interactions between datasets, between datasets and documentation, and between datasets and computing code can result in different types of relations. Some of these fall into “research information” domain, a term used to describe everything associated with research
process: outputs, grants and projects, equipment, researchers and their affiliations, activities, awards, impact, and so on (Bryant et al., 2017). The complexity of research data and information warrants a closer examination of concept relations in order to provide semantic and technical infrastructures for research data management.

This paper uses a framework (Figure 1) to analyze the relations in and between four major types of concepts or entities: data, code, publication, and researcher. The relations will be examined both from the intra-database and between-database perspectives as well as through the lens of association, provenance, discovery, and credit. Association is a generic term used to refer relations between two different concepts or entities, which may be as simple as linking by identifiers, or more complicated ones such as interactive, dependent, causal, or derivative. Provenance refers to relations that will allow researchers to verify and track data and code for quality control and/or reproducibility. Discovery and credit are straightforward and need no further explanation. Each of the four entities in Figure 1 may be examine from these four lens to accurately define the types of relations within and between the entities.

Case studies

Research data in different disciplinary fields vary greatly and it is likely that not all entities in the analysis framework (Figure 1) will be present in all disciplinary fields, for example, the code entity is not relevant in the repository for curating genome sequences. However, the framework provides some direction for how to capture the essence of relations in research data management. Another point to make is that the discovery and credit dimensions of relations are straightforward, thus this paper will focus on association and provenance dimensions in the limited space for discussion. The following two cases are from two disciplinary fields remote to one another.

Case one: GenBank annotations

GenBank is a data repository at the U.S. National Center for Biotechnology Information (NCBI 2017) for curating genetic sequences. An annotation record in
GenBank contains two parts: metadata and the sequence itself. The metadata section has three components: information about the sequence, publications associated with the sequence, and information about sequence data submitter.

As a genetic sequence data repository, GenBank’s original (or raw) data are the base for many other databases to link and reference. We can call the relations between two data repositories “between-database” relations. For example, the BioSample database provides the source (biological materials) from which data stored in GenBank are derived, hence is the “source_of” gene sequences in GenBank, or reversely, sequences in GenBank are “derived_from” biomaterials in the BioSample database. ClinVar is another database at NCBI that collects user-contributed variants of known-genes (Figure 2). As such, sequences in GenBank are the “source_of” variants recorded in ClinVar, while ClinVar contains variants of the genes that are stored in GenBank.

Figure 2: Intra-database and between-database relations in the example of GenBank

Besides between-database relations, GenBank records also have intra-database relations that are mainly between the records and between elements within the same record. It is common in GenBank that a sequence may be updated and resubmitted as a new version, a portion of a gene sequence with a large number of base pairs may be submitted as a separate record, or a record is an assembly of a collection of sequences submitted at different times by different submitters.

The relations at metadata element level requires an ontological approach to model how researchers, data, and publication are related. Relations at this level have been discussed in detail by Qin and Paling (2001) and Zeng and Qin (2016). Although the relations discussed in these publications are not intended for research data, the ways they discuss element level relations can be useful references.

Case two: gravitational wave research information

Gravitational Wave (GW) research is at the scale of “big science” not only in the sense of disciplines and collaboration scales but also in the sense of financial
investment. As a big science field, the complexity GW data creates many challenges for managing them. The first challenge is how to connect different types of data. The raw GW data are generated by laser interferometers as streams, which will then be calibrated and segmented before researchers analyze them. GW data may be categorized and labeled differently among the scientist depending on who are using the data at which stage of research. For example, recent detection data have been sliced by “Events” and “Bulk”, and the input, output, code, documentation, and publications for each event are compiled as one dataset and given a DOI (LIGO 2017). Figure 3 shows a simplified flow of data and information in a typical GW analysis project.

Figure 3: Gravitational Wave research information overview

Note
1. Git_Repository: scientists use GitHub to store their scripts or executables.
2. Configuration: a file in which everything needed: data, parameters, executables, and output location for running the analysis is listed.

It is obvious that entities in Figure 3 are associated with one another in different ways. For example, within the Executables entity, it is common that the execution of one piece of code is dependent on certain software programs or a particular version of the software to work, hence code A is dependent_on code B would be the dependency relation (Figure 1) within the code entity, while Executable is maintained in GitHub and referenced in Configuration.

As GW research is highly data and computationally intensive, provenance relations play a significant role in making sure data, code, and workflows can be tracked all the way to the starting point, should any of them need to be verified for reuse or quality control purposes. The relation types for provenance focus on who designed the configuration, wrote the code, and compiled the workflows, which data segments were selected from where, and where the output was sent. In a research environment where everything is done through writing programming code, such relation types will also
require to be programmable for automatic capture during the whole computational process.

**Discussion and conclusion**

The two (very brief) case studies offer us a number of lessons. First, there is not a single universally applicable relation typology for all disciplines when it comes to RDM. Each discipline has its own idiosyncrasies and needs domain analysis at several levels together with a thorough understanding of the domain as well as scientific and technical requirements for managing the data to enable discovery and use. Second, A relation typology is not a simple task to identify a list of labels or terms, but rather, it is a challenging task requiring a comprehensive set of knowledge and skills and a team of different talents. Relation identification is not only a process of identifying what concepts and entities there are, but also how they are related within the database, across databases, and throughout the whole research lifecycle. Finally, the ability to generalize relation types for a research domain marks how mature and effective the RDM practice is in that domain, because the existence of a relation vocabulary signals the depth of research and development in that domain. The relation vocabulary in UMLS is an excellent example for this point.

What should an RDM relation typology be exactly? The answer is that it depends. The framework proposed in this paper offers a starting point to identify and map out relation vocabularies for RDM, no matter for which domain you are defining and the relations. If we take association, provenance, discovery, and credit as the four dimensions of RDM relation typology, then the levels – cross-element, intra-database, cross-database – will be a way to cross-cut each dimension for getting a comprehensive picture of the relations in that domain. Some relation types can be prescribed for any domain, *e.g.*, the classical relation types such as hierarchical, sibling, part-whole, but many more cannot be prescribed by a universally agreed list of relations in specialized disciplinary fields. The importance of relation typology will only increase as more disciplinary fields become data-intensive.

Relations have been an important area but not a popular topic of KOS research. This paper addresses the need for relation typology study and argues its importance in research data management in particular. The framework for relation type analysis provides some starting point for conducting domain analysis for identify relation types for research data management.
References


Reorganization of knowledge in technical documents for the information system: the extraction of logico-cognitive information from heterogeneous administrative data

Abstract
An administrative document contains instructions to follow or to make according to the views of the producer, and an administrative document can be defined as a technical document "instructor", which specifies the processes to be followed to reach a goal. This article deals with the issue of technical documents produced by French administrative institutions to manage situations of disability. This documentation impacts the organization of the work of the professionals in activity and regulates the classification of the requests of handicap. Based on a linguistic-informational analysis of a body of technical documents (institutional forms), we have developed a logico-terminological model to accompany the input of information for users with disabilities. The contribution of terminological and information-communication processing in the design of an organizational information system is based on the use of textual and graphical data that exist between textual blocks and logical inferences within the technical document. We describe a methodology to delimit the formal frameworks of documents and their exploitation for an optimal reorganization of the information capture, according to the format of the documents of the French Social Affairs Administration (MDPH), in order to propose a structure of information access in accordance with the needs of users with disabilities.

Introduction
A technical document of the MDPH (Departmental House of Persons with Disabilities) is an administrative document containing technical instructions to follow according to the visions of the producer. The mission of the MDPH is to support individuals with disabilities individually and to meet their specific needs. The purpose is to present in a complete chronological way the process of filling a composite file (identity, region, contributing factors, etc.) so that the evaluation team can make a quick decision. This composite file, thus, constituted is a technical document with specific structuring.

This article deals the impact of their organization on their use by practicing professionals and families of applicants in the field of disability. Based on a linguistic-informational analysis of a body of technical documents, we have elaborated a logico-terminological model of support for the collection of information for users with disabilities for the MDPH, according to a cognitive reference-oriented approach or the documentary universe conveys information and communicational operations for reorganization of information.
1. State of the art on technical documentation

1.1. Textual analysis models in technical documents

In the theories of the document whose founders are numerous and in particular Buckland have questioned "what the document" (Buckland 1997). Based on the previous search work, Tricot et al. define the document as follows: "A document is an object with or without inscriptions, which has been conceived as a document and is perceived as such: carries a communicative intention” (Tricot 2016). At its origins, Pêcheux's work focused on technical documents (parliamentary speeches, trade union texts, etc.) or his discourse analysis was based on a multidisciplinary (qualitative and/or quantitative) approach to content (Pêcheux 1969).

The Piveteau report (Piveteau 2014) had already pointed to the lack of anticipation and good management of the project in the life course, often because of the delay of a diagnosis. It also highlights the lack of implementation of even minimal support. Through these elements, the report emphasizes the need to articulate orientation and support. Thus, the greatest need of these people seems to be moving more towards social mediation and long-term accompaniment taking into account the key periods of the person's life: - transition from childhood to adolescence, from adolescence to adulthood and from adults to old age -, etc. This support is based on a constraint that is to fill a technical document aimed at defining the needs of the person with a disability.

1.2. Predicative/terminological approaches for argumentative analysis in statements

We are interested in the organizational analysis of technical documents (forms, procedures, etc.) of interest for content analysis ((Adam 2001), (Kwasnik et al. 2001) and (Buckland 1997), but our methodology is inspired by linguistic methods (Le Guern 1991) and methods logic (Larouk 1994) to propose recommendations to improve textual content and structure (form, order, etc.).

Breakthroughs in formal analysis have come from linguists like Bally, and logicians who have dealt with the problems of knowledge representation with predicative relationships (Larouk 1994). A certain relation of order makes it possible to separate the two constituents of the preaching in theme and in predicate. A document thus contains linguistic inscriptions that convey speech acts. An act of speech is an act of expression and an act of representation, as Charles Bally (Bally 1944) puts it, by focusing on enunciation. A predicate is, what is asserted, about a referent, designated by the subject. The work of Michel Le Guern (The Guern 1991) goes through an update of the term or it specifies that: "The descriptor is not the word of the language as part of the lexicon (language), but the word updated. This update requires the construction of the phrase. The descriptor is not a word of language but a syntagma of speech. "In this sense, the content of the document would be represented by the list of extracted nominal phrase that refer to the extralinguistic discourse”. This approach of Le Guern
is also based on the semiotic theory of C. S. Peirce to determine the referential mechanism of the universe of discourse.

Based on the observations on the textual documents provided by the MDPH, the question arose as to which elements describe a technical document, including which salient elements represent the intra-text discourse targeted by the linguistic content.

2. What logico-cognitive analyzes for the organization of technical documents?

2.1. Specificities of the MDPH corpus

Our corpus consists of technical documents produced by the Social Affairs Administration for the disabled public. The handicapped have motor, physical, auditory, visual, and/or psychological impairments. To improve their social status, and to facilitate their requests, an analysis of the content of the documents becomes necessary in order to propose a simpler process of access to administrative digital information.

Figure 1: Example of a fragment of the technical corpus (Block with title)

We note that the technical documents of the MDPH are composite with linguistic data and graphic data (images, frames, abbreviations, symbols, etc.), which depend on the producer of these paper documents, and intended for applicants. We observe mixed structures composed of linguistic signs (titles, legends, summaries, acronyms, etc.) and iconographic signs (figures, tables, etc.). The linguistic content of these forms is intended for the needs of people with disabilities, but also professionals of health as social workers. We are interested in the support and the content in a perspective of reorganization of the information in order to facilitate the access to the expressions
conveyed in these working documents (institutional or administrative) by observing the technical organization and the procedural contents, via the linguistic methods and logics of textual content analysis.

2.2. Towards thematic and logico-cognitive analyzes of contents

We will delineate the notion of textual content related to the discourse used in technical documents and explore the concept of textual enunciation. We ask ourselves the following questions: What theme are we talking about and how is it delimited in relation to linguistic and logical elements? In our corpus, a document connects a producer (administrator, author), and an applicant (handicapped) via two elements that are a speech (an argumentative and iconographic content) and a support (paper).

-a). Logico-linguistic analysis of content (graphic and textual)

The technical document can be defined as an "instructor" document, which specifies the process to achieve a goal as shown in the following example:

<1> / the handicapped child walks with a stick /
<1>/ l’enfant handicapé marche avec une truelle/

The preceding statement <1> can be represented by the structure of the following binary predicate: VERB (x, y) = PREDICAT (x, y) ----- WALK (x, y)  x WALK y

where X represents "the handicapped child" and the variable Y represents "a stick".

-b) Implicative co-referential analysis

The terminological unit has features in common with the technical vocabulary as in the example:

<2> / Your family caregiver (the person who takes care of you on a daily basis) wants to express his situation and needs, he can complete part F (block F) /
<2> / Votre aidant familial (la personne qui s’occupe de vous au quotidien) souhaite exprimer sa situation et ses besoins, il peut remplir la partie F (bloc F) /

A (a) □ F

The discourse process updates the term in context "Your family caregiver", where explanatory factors of textual structuring (<the person who takes care of you on a daily basis>) contribute to the textual cohesion of the sentence.

<3> / Need help in everyday life, to/for prepare meals or pay expenses (rent, food, etc.) /

(X to Y ) = (X Y)

< 3> /Besoin d’aide dans la vie quotidienne, pour préparer les repas ou régler les dépenses (loyer, alimentation, etc.)/

( X pour Y ) = (X Y)
Relations between the terminological units and the different binding units (A → F) are present in the sentence sequences.

c) Causal and conditional analysis

Logical relationships between terminological blocks result from textual consistency that returns a precise reference pertain to the cognitive and communicative function of the specialty language

<4> / Je demande le renouvellement de mes droits à l’identiques car j’estime que ma situation n’a pas changé.../.

<4> / I request the renewal of my rights to the identical because I feel that my situation has not changed ... ./.

\[ (X \text{ because } Y) = (X \to Y) \]

<5> / If the person over 18 has a protective measure THEN his guardian answers with him or his guardian accompanies him in his request./

<5> / Si la personne de plus de 18 ans a une mesure de protection ALORS son tuteur répond avec elle ou son curateur l’accompagne dans sa demande./

There are conditional formulations (if X, then Y) that express the cause or the consequence, but the descriptions were constructed in such a way as to create a referential continuity between two successive sentences. Each sentence referred to an object that had been mentioned in the previous sentence with anaphoric retakes.

<6> / If the child is not currently in school, he is too young ... ./

<6> / Si l’enfant n’est pas actuellement pas scolarisé, il est trop jeune.../.

Logical statements specify instructions of the type (if A, then B) = \{(A → B) with A as action and its consequence B\}. The discourse universe is generated respecting the structure (if condition A, then action B) in a referential continuity that classifies the requester. Our review of technical administrative data, via our methodological approach, based on a linguistic analysis allows searching the relevant themes. A logical analysis of the discourse conveyed by the texts, by exploring the links between textual elements within the sentence, made it possible to propose several universe of analysis that we present in what follows.

3. Knowledge representation model to guide the discursive analysis within the technical document

3.1. Modeling the universes of the technical document: which logico-cognitive schemas referential of the discourse?

We describe the formal frameworks of the documents and their exploitation for an optimal reorganization of the information capture, according to the format of the documents of the French social affairs administration MDPH
In order to identify specificities in the technical texts to propose a navigation prototype with a thematic tree structure, we propose a model of knowledge representation (three classificatory universes) to guide the discursive analysis. The transversal themes allow intra-documentary navigation between textual blocks. The inter-proposal themes allow to finely indexing the informational content (co-reference, implication, causality, etc.) of a document by its composite structure.

Figure 2. Method of describing the logic-cognitive structure of the referential functions on document

<table>
<thead>
<tr>
<th>Document Universe</th>
<th>Contextual Universe</th>
<th>Logical-Cognitive Universe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Productive universe of administrative documents</td>
<td>Contextual universe of intra-document sequential blocks</td>
<td>Reference universe of the informational discourse of themes and logical relations.</td>
</tr>
<tr>
<td><img src="document_universe_diagram" alt="Document Universe Diagram" /></td>
<td><img src="contextual_universe_diagram" alt="Contextual Universe Diagram" /></td>
<td><img src="logical_cognitive_universe_diagram" alt="Logical-Cognitive Universe Diagram" /></td>
</tr>
</tbody>
</table>

i. Themes - Titles:

\[ T_{1,T}, ..., T_{k,T} \]

ii. Intra-speech NP-themes:

\[ T_{1,NP}, ..., T_{k,NP} \]

iii. Causal / logical relationships between proposals:

\[ P_1 \rightarrow P_2, P_3 \rightarrow P_4; \text{etc.} \]

iv. Implicative relationships between blocks:

\[ B_1 \rightarrow B_2 \rightarrow B_3 \rightarrow \ldots \rightarrow B_K \]

i. Extraction of the title-themes in the form of a list.

ii. Extraction of the referential SN-themes of the speech.

iii. Extraction of themes in causal / logical relations / subordination between proportions.

iv. Extraction of the themes in the implicative relations between blocks (inter-propositions).

Our analytical framework allows the identification of related topics (titles and intra-speech), aggregation with intra-stated and inter-proposal referential themes lead to the logico-cognitive referential patterns of discourse. The inter-proposal themes allowing to finely index the informational content (coreference, implication, causal, etc.) of a document by its composite structure. In order to identify specificities in texts for a
thematic tree structure, we propose a model of representation of knowledge to guide discursive analysis, in the form of structural axes of a block or cross themes allowing the intra-documentary navigation between textual blocks.

Sequencing is related to the relationships of implication, subordination, and causality between-blocks generate themes and associated sub-themes to lead to logico-cognitive regimens referential corpus.

3.2. Proposal for extraction of logico-cognitive knowledge intra-technical document for the qualification of users' status

We are interested in support and content in a reorganization of information perspective to facilitate access to the semantics of expressions conveyed in these working documents (institutional or administrative).

Thus, in view of the analysis carried out, it appears that the information blocks between the forms studied are not always explicit, which leads to blockages or ignorance because the contextual context proposed by the blocks of information generates difficulties related to administrative discourse (documentary universe). Consequently, the logico-cognitive universe is also impacted insofar as it refers to a frame of reference of the medico-social and administrative universe, which is unknown to the applicants. We note that these documents intended for the general public (disabled and families) are designed to be consulted and interpreted in the framework of tagged operations such as filling a file and access to public aid. So, this normed document consists of statements formed through sequences of speech (procedures) produced by the administration. These textual documents reveal informational details that qualify an application with exclusion or inclusion in a specific statute, such as the following propositions:

<6> / If the person concerned is under 18, (his / her) parents are asked to answer for it.

<6> / Si la personne concernée a moins de 18 ans, ses parents sont invités à répondre pour elle./

These argumentative procedures specify the classifications of a disabled person in a particular category, including his or her direct status, linked to the intersection of variables such as age, or grade level, or university level, etc. which determines the status of the applicant, and categorize the beneficiary (disability costs, place of reception (specialized establishment, hospital, school, high school, university), etc. We find in this space, other indices characterizing this son of speech with communication functions of the documents to locate: the territorial demand (urban / suburban), the targeting the state financiers (state / department / metropolis), etc. The aim is to present in a chronological way the process of filling a led file / driven by the status and not by other criteria (region, etc.), so that the evaluation team can make a quick decision,
following the implicative and causal relationships (see § 2.2).

4. Proposal for an information scheme for an administrative decision-making of "user" centered mediations affected by a disability

The encyclopedic dictionary of information and communication sciences (Lamizet and Silem 1997) defines mediation as "a body that ensures in communication and social life, the articulation between the individual dimension of the subject and its singularity and the collective dimension of sociability and social link ". This definition highlights the fact that mediation can play the role of "binder" between an individual and society. It is thus based on the principle of inclusion, but this concept is based on a wide deployment both at the level of the grounds and the various publics. This mediation is also based on a minimum of information sharing on a given situation. In the case of disability, mediation is also based on a minimum of information sharing on a given situation. The mediation process is done by a third party. These intermediaries or third parties seem to have a fundamental role in the process of social inclusion, which leads us to question the identification of social mediation actors. In the case of the medico-social sector, a function is decisive for this inclusion, it is the social worker.

Our proposal follows the diagram above which highlights the mediations optimized for a user when he asks for help. Each mediation configuration chosen by the user refers to a particular strategy and indicates the ability to understand or not interactions with the institution, via digital documents. Sequencing is related to the relationships of involvement and causality observed between blocks generates themes and associated sub-themes to lead to the logico-cognitive reference schemas of the document and thus to characterize the demand. The linguistic contents of these technical documents must be reorganized according to the "status-disabled" logic so that these administrative formalities are well interpreted for care and / or referral to a specialized center.

Documentary mediation contains technical and linguistic terms that generate inadequate support and guidance, which must be reduced by the access filter linked to the "status" of disability.
The methods of mediation that we have proposed can be completed by the institution through the simplified sections and facilitated the seizure to the carer (family, referent, social worker). This mediation scheme filters interactions to access only the relevant information by status.

The recommendations based on our analysis are summarized in the tree structure below, which reproduces the procedures to accompany a user in the search for a dedicated and optimized path. We can immediately distinguish the fallout from such a tree-shaped result for programming the information system refining the information entry process. The challenge is to match this information with that present in the evaluation framework for compensation. However, there is a real gap between the evaluation tools and the forms to be completed, since the questions asked to the (disabled) applicant include technical ambiguities (abbreviations, decoding, etc.) requiring mastery of the administrative language, hence our proposed schema, which is summarized in the figure below, reproducing the procedures to accompany a disabled person in his search for a course adapted and optimized by the qualification of the status of the applicant.

The methods of mediation that we have proposed can be completed by the institution through the simplified sections and facilitated the seizure to the carer (family, referent, social worker). This mediation scheme filters interactions to access only the relevant information by status.

The recommendations based on our analysis are summarized in the tree structure below, which reproduces the procedures to accompany a user in the search for a dedicated and optimized path. We can immediately distinguish the fallout from such a tree-shaped result for programming the information system refining the information entry process. The challenge is to match this information with that present in the evaluation framework for compensation. However, there is a real gap between the evaluation tools and the forms to be completed, since the questions asked to the (disabled) applicant include technical ambiguities (abbreviations, decoding, etc.) requiring mastery of the administrative language, hence our proposed schema, which is summarized in the figure below, reproducing the procedures to accompany a disabled person in his search for a course adapted and optimized by the qualification of the status of the applicant.
Conclusion

The MDPH encounters difficulties in handling technical files for the compensation of persons with disabilities. This work of analysis consisted of extracting salient elements and projecting a mode of reorganization of the documents and we proposed a categorization of the information units contained in the documents, via logico-cognitive elements. This proposal would be appropriate to federate the requests according to a dynamic process in the interaction with the applicant (handicapped and family), via a system of assistance to the finalization of the procedure.

As a conclusion, and after this logico-cognitive formalization based on the referential and info-communication analyses, we proposed our classification by status in order to facilitate access and avoid an inflation of headings. This functional structuring avoids complex combinations of situations where procedural messages are numerous which may discourage the applicant (handicapped person) from completing his application. What can be learned from a document that is too long and tedious to complete? Where the themes are drowned in an administrative mass of information which had to be reorganized the operations of handling the files of the disabled.

References


Philip Hider

The Retrieval power added by subject indexing to bibliographic databases

Abstract

Earlier studies have found that topic queries would lose, on average, about a quarter of hits if bibliographic databases did not incorporate the subject indexing based on controlled vocabularies and typically carried out by information professionals. Topic queries were run on three more databases, namely ERIC, PsycINFO and SocIndex, to investigate whether this loss level might be generalizable. Results indicate that the amount of “retrieval power” lost varies across databases, probably due to a range of factors. In two of the three databases studied here, over a third of hits were lost on average, whereas in the other case about a sixth were lost. These results point, if anything, to an even stronger case for retaining the assigned indexing included in these databases, as well as to the need for further research into the value of professional indexing across contemporary search environments.

Introduction

In a recent study (Hider 2018) the methodology previously employed by Gross et al. (2015) was used by the author to gauge the “retrieval power” of the subject descriptors and identifiers assigned by professional indexers for a particular bibliographic database. Gross et al. (2015) had found that about a quarter of hits would be lost, on average, if keyword searches on a university library catalog were not supported by Library of Congress Subject Headings (LCSH). In the author’s study the mean percentage of hits that would be lost to keyword queries on the Australian Education Index (AEI) was found to match exactly that of the later Gross study, i.e. 27.0%, despite the different type of database and different controlled subject vocabulary (Australian Thesaurus of Education Descriptors). Hider (2018) also reported a second study in which a searcher was asked to use the AEI stripped of its assigned vocabulary, and then to search, for the same topics, on the full AEI, with the subject indexing reinserted. Again, the indexing was found to support the retrieval of a similar proportion, on average across topics, of additional, relevant resources.

The objective of the follow-up study reported here was to gauge the retrieval power added by subject indexing based on controlled vocabularies to other bibliographic databases, as part of a larger program of research examining the search value of professional indexing across a range of search contexts. In order to facilitate comparison with the findings in the initial AEI study, the same measure of retrieval power, and similar methodology, was applied. The remarkably consistent results in the earlier studies, for a university library catalog and a bibliographic database with quite different content, raised the question of whether the “retrieval power” of subject indexing in bibliographic databases and catalogs might be generalizable, to some
extent, in the same way that Lotka’s Law has been found to apply, fairly accurately, to the frequency distribution of name headings in different library catalogs (Smiraglia 2002).

The measure of “retrieval power” can be associated with the experimental design conducted in classical “information retrieval” (IR), in which retrieval effectiveness is measured by calculating recall and precision ratios based on the outputs from individual queries entered into the system. These indexes, and their derivatives, have been used to assess the effectiveness of subject indexing on numerous occasions since they were first developed for the “Cranfield” experiments of the 1960s (Cleverdon 1967), despite the various methodological issues that have been identified their application to real-life search contexts (see e.g. Schamber 1994). Within this positivist framework, the relative value of controlled indexing has been called into question, although the received wisdom is that both derived and controlled indexing have their strengths and weakness, and that their relative effectiveness depends very much on particular search environments (Rowley 1994; Bawden and Robinson 2012). Moreover, they are not necessarily mutually exclusive: controlled indexing, typically created by information professionals, generally adds some value to a database. The questions are how much value and is this sufficient to justify its cost?

Some of the added value that a controlled vocabulary can bring to search is based on what is often termed the “synonym problem” (Beall and Kafadar 2008). Human searchers cannot always be relied on to think of (or spend time thinking of and entering as queries) all the synonyms and related words that might be included in derived indexing on a given topic. If the database is not supported by a search thesaurus – and most are not – this can become a major barrier to recall. Even when searchers do not make use of a controlled vocabulary’s reference structure, and simply enter “keywords”, they may retrieve resources that they would not otherwise retrieve because of the additional subject indexing, without losing any of the resources they retrieve via the derived indexing. These extra resources are likely to be relevant, assuming that searcher and controlled vocabulary use a given term for the same, or similar, concept. The proportion of resources that assigned indexing (typically based on a controlled vocabulary) adds to result sets, across a range of topics, is what this paper labels the indexing’s “retrieval power”.

This measure of the indexing’s value is not quite as refined as the classical measure of recall, as it is not based on “relevance”, but instead serves as a proxy for this measure. Ultimately the number of relevant resources in any result set is a subjective matter, to be determined by the person doing the searching; the subjective nature of “relevance” has been one of the main criticisms levelled at classical “information retrieval” (IR) experimentation, based on recall and precision measures that apply “standardized” relevance judgements (Schamber 1994). (Recall itself is also not a very practical
measure in naturalistic experiments, where the number of relevant resources in a given collection cannot be easily established). Within the positivistic frame of classical IR, however, the “retrieval power” measure is considered, by the author, to provide a reasonable approximation for “recall”, for the purposes of this research: it is likely that the additional resources retrieved via the assigned indexing will, if anything, tend to be more relevant than those retrieved via the derived indexing, as controlled vocabularies are thought to generally provide for more precision than natural language indexing (Gross et al. 2015). Nevertheless, it should be noted that the methodological criticisms that apply to classical IR experimentation also apply to this measure, and are more fully discussed by Hider (2018).

The author is not aware of studies of “retrieval power” in naturalistic settings prior to the research reported by Gross and Taylor (2005) and Gross et al. (2015). Gross et al. (2015) repeated their earlier study on the retrieval power provided by the LCSH on the newer version of the University of Pittsburgh’s library catalog, in which records had been enhanced by tables of contents, summaries, and limited amounts of other “content”. They found that the retrieval power added by LCSH had not, in fact, diminished all that much. Whether similar levels of added retrieval power are provided by assigned indexing across a range of bibliographic databases is not clear, and answering this question would be a first in comparing the added value that professional indexers bring to different databases. The measure may not be so applicable to databases that search on full text, where precision in the higher ranked hits is a more practical consideration, given the very large numbers of resources that are likely to be retrieved, but the reality still is that most bibliographic databases and library catalogs do not retrieve on full text for the most part (as opposed to provide full text, post-search). There are thus many contemporary databases for which the measure is relevant.

**Methods**

The three new bibliographic databases chosen for this follow-up study were: ERIC (https://eric.ed.gov), PsycINFO (www.apa.org/pubs/databases/psycinfo; specifically, the 1987-2017 version) and SocIndex (www.ebsco.com/products/research-databases/socindex). They support different disciplines, and use different controlled subject vocabularies (ERIC Thesaurus, the Thesaurus of Psychological Index Terms, and Sociology Thesaurus). It should be noted that, in addition, the search interfaces used to interrogate the three databases differed in their functionality and indexing rules: for example, words were retrieved on separately in the case of ERIC, but as phrases in the case of PsycINFO and SocIndex. Perhaps most importantly, the SocIndex system provided an “apply related words” option, which resulted in queries being automatically expanded to cover additional words with variant spellings and word forms. The hosts of the databases used in the study were ProQuest, Ovid and EBSCOhost, respectively.

A range of sampling sources was also employed, to best represent what terms might
be searched on these different databases. For ERIC, 55 of the keyword search queries used for the earlier AEI study were re-employed (ERIC covers the same field as AEI, and is international in scope and outlook; the AEI queries might in some cases comprise Australian terminology). For PsycINFO, the “sixty most-frequent multiword terms” identified in a search log from PsycINFO (Yi et al. 2006) were tested, while for SocIndex, the sample of 96 terms was based on lists of “Research Topic Ideas” for anthropology, criminal justice and sociology students found on a webpage of the Frances Willson Thompson Library at the University of Michigan-Flint (http://libguides.umflint.edu/topics/socialscience). Each source has both methodological pros and cons: the sample for ERIC represented actual search queries, but on a different database; the sample for PsycINFO also represented actual search terms, entered on the same database, but was modal rather than random; the sample for SocIndex constituted terms not derived from actual searching, but was thus independent of the database’s controlled vocabulary. The sample sizes were not large, but allowed for indicative findings (with an estimated confidence level of at least 70% for a 4% margin of error).

The terms in the samples were searched for on the relevant database thrice: first, on the full set of keyword indexes; then on all the indexes except for the index with the controlled subject vocabulary; and finally on all the indexes except for the indexes with the controlled vocabulary and the identifiers (additional, uncontrolled terms assigned by the indexers for concepts not represented in the controlled vocabulary). It was hypothesized that the loss of the identifiers would not reduce retrieval power all that much, as they would tend to be terms present in the abstracts, and thus covered in the full search.

Results

Table 1 shows the mean percentage of hits that would be lost per topic if the subject descriptors and identifiers were missing. The percentages for SocIndex are relatively low, but would probably be considerably higher if either the “apply related words” function was turned off, or its thesaurus also utilized the reference structure of the controlled vocabulary. The other two databases, ERIC and PsycINFO, both lost considerably more retrieval power than did AEI, without their assigned indexing. The results, combined with that of the AEI study, suggest that the retrieval power added by assigned indexing varies significantly across databases, and that this is the case even with databases covering the same discipline (such as ERIC and AEI). An interesting finding shown in table 1 is that, contrary to the author’s hypothesis, the loss specifically of identifiers can sometimes make a significant difference to retrieval power. In such cases, the identifiers must have been created more independently of the other bibliographic elements, such as the abstracts. The relevance of the additional resources that the identifiers yield merits investigation.
Table 1: Mean % loss of hits

<table>
<thead>
<tr>
<th></th>
<th>Without descriptors</th>
<th>Without descriptors &amp; identifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERIC</td>
<td>35.0</td>
<td>36.5</td>
</tr>
<tr>
<td>PsycINFO</td>
<td>26.6</td>
<td>35.2</td>
</tr>
<tr>
<td>SocIndex</td>
<td>11.0</td>
<td>17.2</td>
</tr>
</tbody>
</table>

Table 2 shows the median percentage of hits that would be lost per topic, for the three databases. Again, the variation across the databases is considerable, but in all cases the median is less than the mean, indicating losses of much larger magnitude for some topics. Indeed, there are topics with over 50% losses on all three databases.

Table 2: Median percentage loss of hits

<table>
<thead>
<tr>
<th></th>
<th>Without descriptors</th>
<th>Without descriptors &amp; identifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERIC</td>
<td>28.3</td>
<td>30.9</td>
</tr>
<tr>
<td>PsycINFO</td>
<td>18.0</td>
<td>26.8</td>
</tr>
<tr>
<td>SocIndex</td>
<td>9.8</td>
<td>16.4</td>
</tr>
</tbody>
</table>

Finally, the overall percentage loss, combining all the searches, was calculated for each database, and is shown in table 3. There is even more variation here, with the percentages dependent on the relationship between the individual topics’ loss levels and their overall retrieval power.

Table 3: Overall percentage loss

<table>
<thead>
<tr>
<th></th>
<th>Without descriptors</th>
<th>Without descriptors &amp; identifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERIC</td>
<td>85.7</td>
<td>86.1</td>
</tr>
<tr>
<td>PsycINFO</td>
<td>28.9</td>
<td>38.5</td>
</tr>
<tr>
<td>SocIndex</td>
<td>19.5</td>
<td>26.7</td>
</tr>
</tbody>
</table>

Conclusions

Given that the experiments on the ERIC and AEI databases used the same search queries and retrieved on their respective indexes in the same way, the differences in their results suggest that variations in content affect the extent to which assigned indexing increases retrieval power. Differences in both the derived and assigned indexing may well be factors. Nevertheless, the results across the three databases suggest that functionality and the keywords’ source may affect retrieval loss as well. Since discipline influences content and users’ choice of keywords, it can be postulated that discipline is likely to also influence retrieval power. The various possible factors do not preclude an underlying “law” that applies across bibliographic databases, but no evidence for it was found in this study. Rather, the results suggest that any such law would make, at best, a weak impression on retrieval power in the face of other factors that real-life settings involve, both within and without the database. These might
include the size of the database, its data elements, and the amount of indexing, both assigned and derived, the scope of the controlled vocabulary, as well as the disciplinary nature of the indexing language, and the functionality of the system.

Although additional retrieval power produced by assigned indexing seems to vary across databases, even the lowest yields, resulting from systems that poorly utilize controlled vocabulary, appear potentially significant. If searchers are looking for reliability, it is worth noting that some of the losses on SocIndex were much greater than the mean of 16.4%. In fact, 25% or more of hits were lost in the case of almost 30% of the topics. For other databases, retrieval loss would appear to be considerably higher, sometimes more than the loss estimated in the AEI and library catalog studies. Further research, across more databases and in more depth, is needed to assess more fully the extent that subject indexing based on controlled vocabularies adds value to bibliographic databases. The evidence garnered thus far suggests that it is of a magnitude that database providers and subscribers need to consider carefully, before making any decision to cut costs by jettisoning it.

References
Fernanda Gonçalves, Gabriel David

Risks and requirements of an applied Information Science research framework in healthcare

Abstract
The epistemological discussion on Information Science (IS) has contributed to a better understanding of its object of study and to the growing production of applied IS studies in recent years. Many studies highlight the need to strengthen IS interdisciplinary links, alerting to the danger of loss of scientific identity. In order not to be regarded as an applied discipline, dependent on a wider scientific area dedicated to the conceptual nature of information and its dynamics, the evolution of IS demands the validation of concepts, theories and research methods in organizational contexts. The theoretical discussion of IS’s scientific status is not merely academic or philosophical. It has an impact on the organizational perspective concerning information management (IM) functions.

Considering that the scientific validation of IS depends on epistemological and theoretical research, as well as on the results of applied research, this study presents the development of an applied IM research framework, based on the quadrupole research method, designed to explore the impact of IM function monitoring in a healthcare organization. Results focus on the requirements raised during the development of the research proposal, identifying epistemological, theoretical, technical and disclosure aspects of the study. Conclusions show that the scientific characterization of IS applied research (object of study, concepts, theories and method) contribute to strengthen its interdisciplinary nature and to ensure the future recognition of IS contributions in the production of transdisciplinary knowledge.

Introduction
In 2005, in an IM class at the Faculty of Engineering, Porto University, Tom Wilson provoked his students by affirming: “not everything that has science in its name is, in fact, science”. Wilson’s statement was a reaction against another discussion on IS scientific status, which he regarded as useless since the focus should be on results and not on epistemological premises. The discussion was familiar to those who had an archives degree and had noticed that techniques in documentary sciences were being replaced by theoretical and epistemological concerns. To those with an engineering background there was no discussion because applied research is about results that you can experience, measure and module in a prototype to be tested and repeated. Was the IS School in Porto accepting to engage in an interdisciplinary approach, refusing to reduce scientific thought to technical and technological approaches? Could this approach be commonly validated by humanities and engineering and result in effective transdisciplinary research in IS?

Since then, epistemological discussion has been enriched by theoretical, strategic, methodological and technical contributions. The links between IS and other scientific areas have been strengthened and IS scientific recognition has grown. There is a demand for information professionals in organizations, demonstrating the link between
academic knowledge and the professional market. Nevertheless, many IS research studies lack scientific identity, because results are limited to epistemological or theoretical reflection or due to the lack of application of IS approaches to the study of information. Moreover, scientific identification is not about the topic of the information studied but about the perspective towards the information studied and how you study it.

These concerns constitute challenges for IS researchers and were carefully addressed in the planning of an applied IS research proposal. Which scientific and epistemological concerns should be addressed and to what level? Besides IS theories and models are there any IS technical approaches or models that may be useful? How can these interdisciplinary contributions be integrated in IS applied research without loss of scientific identity? Since the research project aimed to address the relevance of IM functions in a hospital, there was a concern regarding methods and approaches that could best suit the study of IM activities in a healthcare context.

The work includes a reflection based on IS literature regarding the main changes that have occurred in IS and on the characterization of information as an object of study. Discussion and results focus on the development of the scientific framework for applied IS research in the healthcare sector and identify the theoretical, methodological and technical premises which contributed to the critical analysis and development of the research proposal.

1. From epistemological reflection to applied IS

The literature on the evolution of IS or library and information science (LIS) focuses on technical and professional aspects, as well as the scientific status of IS (Adriaenssen and Johannessen 2016; Bates 1999; Hjørland 2014; Saracevic 2009). The fragilities of the IS scientific field demand critical thinking in the development of a research project, not only due to the multidisciplinarity of its object of study but also because the research proposal must be sustained upon a broader interpretative interdisciplinary basis, considering IS connections with other knowledge areas. Although this perspective has been explored in the literature for several years by various authors, like Harmon (1971), the dynamics, evolution and multidisciplinarity of IS still leave these issues open for discussion.

IS evolution dynamics has been explored with the contribution of authors like Silva and Ribeiro (2012) and Capurro (2003), leading to the identification of paradigmatic transition processes. In the Portuguese literature, Ribeiro and Malheiro have identified evolutorial phases in the study of information, based upon the evolution of the informational process in archives: a custodial, historical, patrimonial and technical paradigm (18th – mid-20th century) and a post-custodial, informational and scientific paradigm (from the mid-20th century) (figure 1) (Ribeiro and Silva, 2016; Silva,
Ribeiro, Ramos and Real, 1998). The same development is characterized by Capurro (2003), in a perspective focused on information retrieval and the introduction of technological systems, identifying three epistemological paradigms in IS: physical, cognitive and social (figure 2).

Figure 8: IS paradigms according to Silva and Ribeiro

<table>
<thead>
<tr>
<th>HISTORICAL, CUSTODIAL, PATRIMONIAL AND TECHNICAL (between the 18th and the mid 20th century)</th>
<th>FOCUS: HISTORY AND ACCESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nation-Archive concept; document incorporation politics and reclassification of documents</td>
<td>Historical value</td>
</tr>
<tr>
<td>Valuation of uncertainty, preservation and restoration functions</td>
<td>Dominance of the technical perspective</td>
</tr>
<tr>
<td>Growing importance of the access to the documents content led to the creation of retrieval instruments</td>
<td>Information retrieval and management (IRMA)</td>
</tr>
<tr>
<td>Investment in classification and indexing models (Development of Documentary Sciences to work with scientific technical information) (19th and 20th century)</td>
<td>Didactic teaching/learning for archivists and librarians</td>
</tr>
<tr>
<td>Theoretical production concept</td>
<td>Records group and records management (USA)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>POST CUSTODIAL, INFORMATIONAL AND SCIENTIFIC (from the middle of the 20th century)</th>
<th>FOCUS: INFORMATION AS OBJECT OF SCIENTIFIC STUDY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information understood as a human and social phenomenon</td>
<td>Information dynamics vs. documentary</td>
</tr>
<tr>
<td>Information access as guarantee for transparency</td>
<td>Changes in the theoretical and professional framework</td>
</tr>
<tr>
<td>Adoption of scientific theories and models to understand social information</td>
<td>Technical perspective is substituted by a scientific logic</td>
</tr>
</tbody>
</table>

Adapted from Ribeiro and Silva (2016)

Ribeiro and Silva (2016) claim that Capurro’s paradigms are approaches that represent a fragmented disciplinary perspective of IS, but both highlight the multi and interdisciplinarity of IS and the wide reach of its research field, just as earlier pointed out by Borko (1968). The authors stress the theoretical complexity of the study of information and the value of IS studies in organizational contexts. The evolution of IS from documentary sciences to a scientific paradigm should be regarded as a favourable opportunity to validate IS’s status through applied research. Nevertheless, there is still some resistance and criticism among archivists and librarians, fearing that the scientific approach contributes to a disregard for the value of the technical work of information representation and retrieval. The discussion of the scientific requirements of IS studies demands critical thinking in the development of applied IS research, contributing to strengthen IS’s scientific status and demonstrating the value of IS professionals’ competencies. The true challenge is to understand whether interdisciplinary applied research, focused on results, will evolve to produce transdisciplinary knowledge, or whether the influence of technological developments will contribute to reduce scientific thinking in the field to technical approaches and technology, as Helena Santos (2014, p. 29) points out.
Figure 9: IS paradigms according to Capurro

Adapted from Capurro (2003) and Almeida; Antonio; Boccato; Gonçalves and Ramalho (2007)

The complexity of information systems and the dynamics of an information society have enlightened the interaction between the production, storage and reuse of information, reinforcing connections between various knowledge areas. The interdisciplinary contributions in IS research, purely centred on epistemological discussion, signify a lack of focus on scientific production. As a result, this could contribute to the isolation of IS and a weakening of scientific interdisciplinarity. In this context, it is important to note Floridi’s (2002, 2004) view of the connections between Philosophy of Information and IS. The integration of IS into a much wider scientific area dedicated to the conceptual nature of information and its dynamics may lead IS to be regarded as an applied discipline. IS scientific production must be independent of the object of study and results must be concrete (Floridi 2002, 2004). This does not mean that epistemological, theoretical and methodological concerns are not essential to applied research and its scientific validation. The evolution of IS does not require an exhaustive literature review on its scientific status but guidance as to critical thinking on the development of research studies.

2. Critical thinking on the development of applied IS research

There are various scientific methods, already validated in other areas, that may be used in IS research. The question is not the number of IS research studies produced but
the identification of those studies with IS as a scientific area, and the impact and results of such studies.

Implementing results of IM studies in organizations demands systemic organizational changes, involving socio-cultural and socio-economic considerations which constitute barriers to strengthening the importance of IM as a strategic organizational function. Although IM is recognized as strategic, and information resources are critical to organization, the lack of a formal organizational IM strategy contributes to focusing IM decision making on technological infrastructure development, as other IM issues are addressed at an operational management level. As a result, much IM research ends up adding to the lack of consistency between information infrastructure and the processes of information production and use. Since IS scientific validation depends on epistemological and theoretical research, and on the results of applied research in IS, researchers must invest in developing robust and solid research frameworks in order to strengthen its scientific position.

Since the process of identifying and valuing certain features in an information context, so that a study can be reproduced in other contexts, is highly dependent on the technical approaches used, applied IS research demands critical reflection concerning the approaches, models and strategies used in the research process. The characterization of IS research identified by Silva and Ribeiro (2012) is a reaction to the purely technical approach of the archival discipline. The production of a technical work of information representation or retrieval, such as an online catalogue, has value as an information product. Validation of IS research depends on the method, concepts, theories and technical strategies used in the process to support its development, enabling a replication of the process in another context. The researchers’ work is to identify IS premises, concepts, theories and technical approaches that are valuable to the study of the information object, considering its context of study as well as the results defined. These premises must be critically identified in the research proposal development and enhanced during the research process, connecting epistemological reflection, theoretical guidance, technical approaches and instruments to be used, and the impact of the study’s deliverables and results.

The use of IS concepts is important for consolidating IS. Using an information unit instead of a documentary unit, for instance, does not need to be limited to the physical format of the information. Conceptualizing archives as information systems is not synonymous with the archival description “archive facility”. The concept encompasses all the organization and actors responsible for producing, reusing and managing information, within an integrated and dynamic approach to information. The application of a systemic approach to the study of information results in an integrated view of informational organization, considering the different elements that constitute information and how it is organized and structured (informational structure), as well as
the processes of interaction between all its elements and other external systems (dynamics of informational processes). In this regard, the organic-functional study constitutes a basic conceptualization scheme of informational structure to be combined with the adoption of the systemic approach to the study of information, which implies a dynamic perspective of the information system, with its inputs and outputs, considering the interdependency of its components and their interaction with other systems. When affirming that the research framework in the informational paradigm allows the study of information as a phenomenon and process, Silva and Ribeiro (2002) use Morin's theory of complexity. The adoption of the systemic approach reinforces interdisciplinarity, expanding knowledge to apply to the complexity of organizational phenomena, considering the dynamism of its functioning and its relationships and interdependencies with the environment. In the research process, this interdisciplinarity may be supported by using the quadrupole method, identifying the concepts, theories and methodological approaches to the operationalization of research in a dynamic process that allows us to overcome disciplinary isolation and the distortion of knowledge, and to comprehend the complexity of the system itself, linking inputs and outputs.

3. Development of an IS research proposal in a healthcare context

The conception of the present research project began in the context of organizing an information management (IM) service for a Portuguese hospital and has its foundation in previous IS academic research (Gonçalves 2011) (Gonçalves and David, 2013). The initial focus was on research questions, the hypothesis that could be stated and the significance of the contributions to the study. The technical, methodological, theoretical and epistemological premises of the study were only discussed after its focus and aims had been clearly defined.

Since we intend to study the contribution of IM in a hospital context, we asked the following research question: How can monitoring information activities, based on performance indicators, contribute to improving healthcare information management? To understand the impact of IM function in healthcare activity, (objective) results of the study focused on: (i) the identification of performance indicators that characterize the information model of an hospital; (ii) the development of a monitoring model for information management in the context of the clinical activity of a hospital and (iii) the application of the monitoring model and assessment of its contribution to supporting IM decision making and its operationalization.

The study demands an analysis and explanation of mechanisms of clinical information production and use, which must be supported by a mapping of the information processes. Moreover, it is essential to characterize the information infrastructure of the hospital, to support the production and use of clinical information, and gain an understanding of its lifecycle. Although, these two aspects are an essential
part of the study and will enable the production of information products that are important to support IM activities, as well as help develop the monitoring model, it is also essential to identify expected results so as to be able to precisely define the research contribution.

The theoretical and methodological concerns identified as crucial in the initial stages of defining the research were: (i) the identification of IS concepts, theories and technical approaches that were considered worth applying in the study; (ii) the adoption of a scientific research methodology to ensure the scientific validation of the study and (iii) the adoption and integration of theories, methods and technical approaches from other knowledge areas, characterizing the interdisciplinary dynamics of their use in IS research. The research question, the aims of the study and definition of its results were followed by a literature review, focused on IS and interdisciplinary models and approaches that could be important for characterizing and studying an informational context, considering the relevance of its application in a hospital context.

Two particular aspects increased the need for a methodological framework. The fact that the development of the research proposal led to the application of interdisciplinary methods and approaches demanded a clarification of the epistemological and theoretical positioning of the study. Simultaneously, it was important to recognize that concerns about the impact of the research process and its results on the healthcare context of the study should not constitute its main aims. Applied IS research in organizational healthcare contexts is easily misunderstood with organization management, computer science and communication areas, among others. The development and application of an adequate research framework was essential for conducting critical thinking and reflection that addressed the results of the study as well as epistemological, theoretical, and technical concerns. In this context, the use of the quadrupole method contributed to strengthening the scientific identification of the applied research.

For a clear understanding of how application of the quadrupole method supported the development of the research framework, results of the research proposal are represented in a graphical representation in figure 4, considering the dynamic relations established between the four poles in the research strategy, with respect to: (i) scientific positioning (epistemology), (ii) theoretical support (theory); (iii) operationalization of the research (technical) and (iv) results (disclosure). In terms of concepts and theories, adopted a priori in the definition of the research strategy, we have pointed out three central aspects that we consider fundamental for the development of the present research problem:

- the concept of information as an object of study, regardless of its application context, considering construction, communication, storage and use of information as informational activities and the properties of information (Le Coadic 1997)
the study of information considering the informational act as a base informational unit, (Ribeiro 1998, p. 28-36), to be considered in the mapping of information processes;

- the adoption of the systemic approach to the study of information, conceiving the informational context of the study as an information system, according to the principles of the General Systems Theory (GST) (von Bertalanffy 1968) to describe how the different components of the organization interact;

- the study of the informational context as a socio-technical system, following the GST and the concept of socio-technical systems developed by Emerit and Trist regarding the adaptations of humans to the organizational and technical component of work contexts Ropohl (1999); Mele, Pels, and Polese (2010), noting the interaction between people (human behavior) and technology and environment (infrastructure) in the analysis and characterization of the informational context, considering transactions that occur within the system and its subsystems and between external contexts in its dynamics with the environment.

As depicted in figure 3, the analysis of the context of information production and use includes an information behaviour approach, as well as identification, mapping and assessment of informational processes and activities using process modelling techniques, for instance. These aspects are identified as theoretical models and data...
analysis techniques, represented in a dynamic framework that shows the impact in terms of scientific positioning, knowledge areas involved and interdisciplinary contributions. The interdisciplinary contributions from areas like computer science and organization management, which are valuable to the research process, including, for example, the use of soft systems methodology (SSM) or balanced scorecard (BSC), are integrated into the research strategy, considering the impact of its use in the epistemological pole of the research.

Conclusions

The research framework based on the quadrupole method presented reflects the initial premises, which will surely be enhanced during the research process. The framework development led to a systematization of critical steps in the development of the applied IS research project that is currently being developed. The preliminary results may contribute to strengthen the interdisciplinary nature of IS applied research, minimizing the risk of loss of IS identity and ensuring the future recognition of IS contributions in the production of transdisciplinary knowledge.

Although the use of the quadrupole method is quite common in IS research, it is not usual to represent \textit{a priori} the relations that will be established during research between the four poles, an approach which is usually used in natural science models. The use of the quadrupole method to identify the dynamic connections between the different poles enables a clear understanding of these relations and may contribute to strengthening interdisciplinary IS research.

References


Roger de Miranda Guedes, Maria Aparecida Moura

Semantic warrant, cultural hospitality and knowledge representation in multicultural contexts: experiments with the use of the EuroVoc and UNBIS thesauri

Abstract
Cultural warrant is an aspect of the principle of semantic warrant and recommends that knowledge organization systems (KOS) should be designed from the premises and concerns of a given culture. Although there is no consensus as to a better means of systematizing or apprehending cultural warrant, it is possible to identify certain traits in KOS that signal their readiness to give more sensitive consideration to cultural manifestations in information recognition. In this paper we have analyzed two terms, namely migration and terrorism, contained in two controlled vocabularies that are known to represent the multiculturalism of their user communities, the EuroVoc and the UNBIS thesauri. The study establishes that the conceptual relationships described by the abovementioned terms reveal coverage of user requirements and attention to the socio-cultural outlook of user communities. The perspectives of both cultural assurance and cultural hospitality were recognized as KO concepts that can account for (or justify) the semantic adjustments necessary to meet user needs in a multicultural context.

1. Introduction
Knowledge organization systems (KOS)\textsuperscript{1} are interpretations of fields of knowledge materialized in symbolic representations intelligible to human beings and machines. They influence how their users interact with information within the KOS itself and beyond. Considering the instruments that represent knowledge, such a statement evokes the complex relations of meaning and information established in the field of knowledge organization (KO).

Knowledge and language are entwined, since language is the symbolic shaping of knowledge in a communicative manner and because language is the instrument through which knowledge is conveyed. In the present paper, language is the starting point for discussion of theoretical and methodological elements underpinning KOS, situated within technological realities that demand semantic strength capable of providing for multicultural scenarios and contexts. In this way, the principle of semantic warrant comes into play – particularly one aspect of it, namely that of cultural warrant – reflecting upon meanings in documentation language set in information contexts that lead to the coexistence of multiple points of view and interactions with the same object of information or to a specific field of knowledge.

\textsuperscript{1} The term knowledge organization systems (KOS) covers all types of schemes, tools and devices that present the organized interpretation of knowledge structures. KOS include classification systems that organize materials at a general level (such as bibliographical classifications), subject headings and thesauri, which allow for more thorough access, and authority control schemes that regulate terminological variation. They also includes less traditional schemes such as semantic networks and ontologies (Hodge 2000; Zeng and Chan 2004).
To illustrate this, we conducted a brief analytical experiment on two well-known controlled vocabularies, namely the EuroVoc and UNBIS thesauri. As KOS, both are oriented towards broad information organization purposes and serve large, diverse user communities with strikingly multicultural natures, presenting challenges to the representation of significant aspects of the concepts relevant to these communities.

2. The Principle of semantic warrant and its perspectives

According to Campbell (2008), the concept of warrant in relation to library science has been a stepping stone towards the understanding and development of systems designed to represent knowledge, providing an intellectual basis for validation and inclusion of semantic entities in these systems. In the author's understanding, warrant articulation in KO is the combination of precepts for the deliberate construction of semantic consensus, using both formal and informal criteria.

The challenge of exploring the principle of warrant began with British librarian E. Wyndham Hulme (1859-1954), who called into question the need for methods and rules in the definition of concepts in a particular field of knowledge. Hulme questioned what could be considered the necessary warrant for determining the set of terms that would reproduce the conceptual domain of fields of knowledge.

One of the best-known definitions of the (semantic) warrant concept in Library and Information Science (LIS) is that of Clare Beghtol, who describes it as

[...] the authority a classificationist invokes first to justify and subsequently to verify decisions about what classes/concepts to include in the system, in what order classes/concepts should appear in the schedules, what units classes/concepts are divided into, how far subdivision should proceed, how much and where synthesis is available, whether citation orders are static or variable [...] The semantic warrant of a system thus provides the principal authorization for supposing that some class or concept or notational device will be helpful and meaningful to classifiers and ultimately to the users of documents (Beghtol 1986, 110-111).

Warrant, therefore, would be the ontological imperative that assures terminological units incorporated into a KOS carry the adequate meaning for the purpose and utility of such device. From the point of view of KO, the relation between meaning and pragmatics – implicit in the concept of warrant – could be attested by different areas of study such as linguistics, sociology, semiotics and philosophy. What connects these fields is language as the intelligible structure through which reality can be accessed.

To Mai (2011), the idea behind warrant is that semantic justification in the inclusion of terms and classes is based on parameters that are external to the beliefs and knowledge of the professionals that develop, manage and make use of KOS. Professionals that represent and classify information resources use such parameters as reference in the conceptualization of a KOS.

To meet the needs of expressing meaning that is useful to users, developers and maintainers of KOS must make informed choices regarding which perspectives should
be prioritized in these systems. According to Howarth and Jansen (2014), it is possible to identify numerous warrants among the development processes of classifying systems and their relationship with users. Several classic warrants, starting with the literary warrant proposed by Hulme (1911) at the beginning of the last century, have been used to guide the creation of classifications.

In the list of warrants proposed in the field of KO, there is a wide range of warrants for the development of KOS such as literary warrant, user warrant and organizational warrant – recommended by the North-American standard ANSI/NISO Z39.19-2005 (National Information Standards Organization 2005) for the construction and management of controlled vocabularies. Also present are warrants that explore "the possibilities of balance between one another so as to assure a methodological support that achieves a representative conceptual structure as a result […]" (Barité 2011, 4), such as philosophical warrant and educational warrant (Beghtol 1986), structural warrant (Svenonius 2000), cultural warrant (Lee 1976), and phenomenological warrant (Ward 2000). Warrants related to specific and/or contemporary needs can also be seen as a necessity when trying to explain phenomena caused by the digital culture – specifically the phenomenon of occupying virtual platforms and environments.

It is worth noting that Beghtol (1986) recognizes cultural warrant as an umbrella term under which all types of warrant can be understood as detailed case studies of that concept, while Barité (2011, 4) sees other warrants as "derived from or part of literary warrant", while also acknowledging the polyvalence of cultural warrant.

3. Warrant and cultural hospitality

In discussing ethical imperatives in the organization and representation of knowledge, Beghtol (1986; 2002) refers to the concept of "cultural warrant", which suggests that any strategy for the organization of knowledge or knowledge representation systems must reflect the assumptions, values and biases of the culture(s) in which it is inserted. The concept of cultural warrant was proposed by Joel Lee (1976) and was inspired by Hulme (1991)’s term "literary warrant". Cultural warrant means that any kind of knowledge organization and/or representation system can only be fully helpful and useful to people of a specific culture if based on assumptions, values and biases of that same culture.

From this standpoint, KOS – being manmade products – are no more than a reflection of societies and their shared values. Cultural warrant is based on the idea that every knowledge organization system represents and imposes a worldview (Beghtol 2002) and is therefore considered a political and cultural artifact (Grolier 1982). These instruments, although not necessarily tangible and, as such, not entirely apparent to the user community making use of them, have a major influence on the individuals of said community.
The corollary to cultural warrant is the concept of hospitality which has been used within the field of knowledge organization to refer to the ability of a classification system to accept new concepts and relations between existing ones. Beghtol (2002; 2005) expands this traditional view of hospitality into the concept of cultural hospitality, in which knowledge organization systems may allow for personal and community choices.

According to Beghtol (2002), the specialized literature considers hospitality the ability of a classification notation to incorporate new concepts and establish semantic and syntactic relations between old and new concepts. Such a definition seems particularly relevant to the treatment of questions about meaning in knowledge organization systems and methodologies pertaining to dynamic environments and multidisciplinary fields.

Cultural hospitality would then be a precept for the creation, development, management, review and use of representation and organization systems concerned with the diversity of meanings in usage contexts that are not limited to well-defined sociodemographic profiles.

Within the concept of cultural hospitality lies an ethical underpinning that creates the conditions for individuals to express themselves and act individually or as a whole while maintaining the possibility of sharing local and individual knowledge, giving way to discussions over particularities and over their own cultures (Boccato and Biscalchin 2014).

Thus, it is evident that the notion of cultural hospitality, despite its potential to substantiate the need for inclusion and expansion of the scope of meanings in fields of knowledge, calls for extensive reflection and careful consideration, starting with the question of what exactly constitutes a hospitable KOS (Guedes 2016).

4. Overview and experimenting on the thesauri

Despite the lack of a practical procedure to quantify the hospitality of a KOS when faced with the contexts it serves, it is possible to draw comparable inferences by analyzing two controlled vocabularies known for representing the multiculturalism of the user communities that make use of them, namely the EuroVoc and the UNBIS thesauri. EuroVoc is a multilingual and multidisciplinary thesaurus for the activities of the European Union, available in twenty-three languages of member countries. The UNBIS thesaurus is a multilingual vocabulary developed by the Dag Hammarskjöld Library (DHL), connected to the United Nations (UN) Department of Public Information, and available in all the official languages of the UN.

Both instruments were designed to answer the information needs of remarkably heterogeneous socio-cultural groups, that being the main reason behind their selection for the experiment. This is a challenge faced by the KOS that must deal with various
criteria that may guide the verification of meaning in its conceptual structures; that is, there is a need to invoke multiple perspectives of semantic warrant that cover levels of meaning increasingly more refined and politically defined.

For a brief empirical analysis of the two instruments of representation, we chose the terms *migration* and *terrorism* in English. The reason behind this choice and the concepts represented by them is due to the active use of both terms in recent years as a consequence of sociopolitical phenomena on a global scale.

We tried to examine the terms *migration* and *terrorism* in both thesauri considering all terms related in some way to them. The aim was to draw comparisons and highlight the presumed attention given by both instruments to questions regarding the manifestation of their audiences’ cultural identity. Figure 1 and Figure 2 illustrate the terms *migration* and *terrorism*, respectively, in both KOS, and their relation to other terms.

![Figure 1: Migration in EuroVoc and UNBIS Thesauri](image1)

![Figure 2: Terrorism in EuroVoc and UNBIS Thesauri](image2)

Source: Adapted from EuroVoc (2018) and UNBIS Thesauri (2018)

We observed that in the case of the term *migration*, the relations that result from branching off into more specific terminology in the EuroVoc thesaurus reveal with respect to the target audience – and to its discourse as crystallized in documents produced within the European Community – a need to particularize notions and concepts arising from facts that have taken place in the tangible social reality of the user community served by the KOS. The same exclusivity in terminological branching does not seem to occur in the UNBIS thesaurus, which shows less specific terms related to *migration*. 
With the historical and social subject-matter of migration flows within the European continent in mind, it is possible to infer that the context of its creation and the audience served by EuroVoc had a higher demand for detailed socio-cultural aspects experienced by that group. This way, the conditions of meaningfulness of said KOS can be contributed to by using cultural warrant as a criterion for representativity. Culture, while seen as a sensitizing element, supports the analytical apprehension of useful meanings for individuals that – while immersed in a game of language – find meaning by relating the term *migration* with terms such as *diaspora* and *braindrain*, for instance.

In the second example, the term *terrorism* is represented in both thesauri, showing its semantic relation to other specific terms. We observed that in the UNBIS thesaurus there is a greater need to represent concepts that are characteristic of a certain social reality. This does not occur with EuroVoc, which presents less links between *terrorism* and specific terms directly related to it. In this case, the influence exercised by the purpose of the UNBIS thesaurus operating on the institutional grounds of the UN, despite its position as an international organism, exhibits an extraterritorial nature that results in a range of domains of activity which, in some cases, require more nuances.

Therefore, while these instruments aim to serve the needs of knowledge representation to their user communities, negotiations of meaning are felt with each decision, betraying thoroughness or negligence regarding a concept. The above example from the UNBIS thesaurus shows the clear need to directly correlate the concept of terrorism with the historical fact of the terrorist attacks of September 11, 2001 in the United States (*terrorism attacks (11 September 2001)*). In this case, the use of the term *terrorism* and its related terminology acknowledges the manifestation of a cultural hospitality activated to answer the need for meaning elicited by historical, social and cultural contexts.

5. Considerations

The diversity of units of analysis in the organization and representation of knowledge – in relation to elements that impact meaning, such as cultural components – have been recognized as defining characteristics of KOS that aim to be sufficiently hospitable to the experiences of user populations on a global scale. It is all due to the greater opportunity for interaction and access to information promoted by digital technologies, enabling the establishment of information environments in which the multiplicity of watchful eyes demands constant updates and refinements from the systems responsible for organizing and representing knowledge.

This study thus acknowledges the perspective of cultural warrant as a concept that comprehends and justifies changes and variations that other perspectives may present due to demands for information or knowledge production by the user community being served. We also highlighted the analytical concepts of KO that are able to explain (or
justify) the semantic adjustment necessary to answer user needs, such as the subprinciple of cultural hospitality, proposed as a way to accommodate in the KOS meanings that serve a diverse and fragmented audience. Devoting attention to cultural warrant and to cultural hospitality is a way of managing the cultural bias that could compromise the global usefulness of knowledge organization and representation schemes, particularly of KOS that meet the demands of virtual contexts.

References


Athena Salaba, Joseph T. Tennis

Solid Foundations and some secondary assumptions in the design of bibliographic metadata: toward a typology of complementary uses of metadata

Abstract
Traditionally, studies on the functions of the catalog have focused on the tasks performed by the users of catalog to access bibliographic information with the ultimate goal to access the content of the resources represented by the bibliographic data. This paper examines the uses of bibliographic metadata by a variety of user types and presents findings on the purpose, functions, and use of bibliographic data to include additional uses by institutions or individuals.

Introduction
Historically catalogs of library resources have taken different forms, the bibliographic data included in a library catalog entry has increasingly been enhanced, and the way users interact with the catalogs and its bibliographic data has also changed, especially since the end of the last century. Bibliographic data include information that describes and represents a resource. Throughout their history, library catalogs have served a variety of purposes. The majority of the literature on the purposes of the catalog focuses on the finding and selecting functions. Cutter, in his Rules for a Dictionary Catalog, identifies three catalog objectives: to find a resource, to show (collocate) what resources a library has, and to assist in the choice of a resource (Cutter 1876). Theories of bibliographic control identify five main catalog functions: find, collocate, identify, select, and obtain, with the additional function of navigation included in fewer works (Svenonius 2000). More recently, the IFLA Library Reference Model (IFLA LRM), a consolidation of the three models, defined the find, identify, select, obtain, and explore user tasks when interacting with the information accessible through a library catalog (IFLA 2017).

Although these have been widely-discussed functions, they seem to only address the uses of the catalog by library patrons and intermediaries who assist them with these user tasks. Often, users of the bibliographic data included in a catalog or other bibliographic data sets are the institutional users, such as staff and administration of a library who use bibliographic data for purposes related to their areas of responsibility. These responsibilities may include cataloging and metadata creation, acquisitions and collection development and management, resource sharing, reference, and various assessment efforts, including institutional accreditation. Most often, these functions of a catalog are referred to as the inventory and management functions of bibliographic data.
This paper examines the uses of bibliographic metadata by a variety of user types. More specifically, this paper presents findings on the purpose, functions, and use of bibliographic data.

To undertake this task we analyzed the literature, focusing on the functions of the catalog, types of metadata to fulfill these functions, and user tasks related to both catalog function and bibliographic metadata. The focus is on identifying bibliographic metadata that supports institutional purposes like collection building, collection quality assessment, marketing, outreach, and educational purposes.

We take a comparative approach to this examination, asking what is particular to user-focused tasks versus institutional uses or requirements. Further, the universe of metadata is larger than traditional bibliographic data, and this fact influences how we conceptualize the universe possible of functions. To that end, we examine metadata standards that support similar information systems, but have a different remit than library catalogs.

**Functions, purposes, or objectives of the catalog**

Others have tried to identify purposes and functions of a catalog taking a different or broader approach. The first to write about the functions of a library catalog was Cutter (1876), who identified the objects of the catalog. Wallace (1984) provides a historical account of the purpose of the catalog. Svenonius, in her 1992 work “Bibliographic Entities and their Uses,” discusses the “functions supported by a library’s file, i.e., catalogue” (p. 10). Lee and Lan (2009) make a distinction between the purpose of the catalog and the bibliographic objectives of the catalog, stating that “bibliographic objectives establish an operational framework for a library catalog” and that “bibliographic objectives are dependent on the catalog’s purpose or purposes” (p. 221).

Tennis (2006), uses a framework analysis to examine the purposes and function of information organization, many of which relate to functions of the bibliographic data included in an information system such as a library catalog. Clarke (2014) presents her study of the functions, goals, and objectives, which she uses interchangeably to mean the ends that library catalogs are or are expected to be designed to meet.

The literature identifies the following purposes of a catalog: to serve as a tool to locate resources, to organize knowledge, to serve as a guide to literature or a particular collection, to establish authority or canonize, to serve as a reference tool, to provide a historic record, to arrange information, and to provide a knowledge base.

Though purpose and function are often used synonymously in the literature, we treat them differently. Purpose, by our lights, is the reason or motivation for why some information system is created. Functions are the actions taken to fulfill the purpose (Tennis 2006). Using this lens, we inventory both the stated purposes and functions in
the literature, build out a comparative typology, and link this to bibliographic metadata used in institutional work like conducting inventories, collection management and assessment.

**Using bibliographic metadata**

Many works focus on the identification of the functions of the catalog but there are other works in the literature that validate functions identified in previous studies or identify additional uses of bibliographic data in an effort to assess institutional research, services, or collections. In addition, the literature offers examples of works whose main goals are not the identification of catalog functions but assessing information systems or designing systems that facilitate user tasks and user needs. A distinction is made here between use of bibliographic metadata and interface functionality of a retrieval system, which sometimes is also referred to as a function of a catalog (e.g., downloading results). Fewer works on the functions of the catalog define the users or base their identification of the functions on user studies, including user information seeking behavior. Cossham (2013) suggests that looking into information seeking behavior (ISB) and catalog use may lead to an expanded identification of functions based on how users use it.

In our analysis of functions, we identified two main types: primary and auxiliary functions or uses of bibliographic data. We define primary functions the actions taken or uses the bibliographic were designed to support. These may often be described as “traditional” functions. Auxiliary functions or uses are other ways the data used to support institutional or individual needs but these are not figured in the design of bibliographic data. In addition, categories of functionality are defined. Table 1 offers an overview of the resulting analysis, including categorization, terms used in the literature for each function, and a definition. Following the Table is a brief discussion of the categories and functions.

Beyond the identification of primary and auxiliary functions, our analysis shows that functions facilitate actions that can be characterized as:

- **access**: actions facilitating access to entities and the content of resources
- **discovery**: actions facilitating the exploration and discovery such as relationships, overview of collections or domains
- **education**: actions that facilitate user instruction on how to search or use the bibliographic data
- **management**: actions that facilitate the management of collections or resources
- **support**: actions that facilitate the use of bibliographic data in support of institutional or individual curatorial, assessment, or other necessities.

In our approach, when necessary as determined by the definition of the function, we have grouped together some sub-functions or functions discussed either as one or as
two separate ones in the literature.

Table 1: Functions and uses of bibliographic data

<table>
<thead>
<tr>
<th>Type</th>
<th>Category</th>
<th>Function</th>
<th>Terms</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>primary</td>
<td>access</td>
<td>finding</td>
<td>find, finding aid, finding, locate, collocate, search, retrieve</td>
<td>to find or locate entities based on certain criteria</td>
</tr>
<tr>
<td>primary</td>
<td>access</td>
<td>identification</td>
<td>identify, differentiate, verify, evaluate, distinguish, make relevance decision</td>
<td>to confirm the nature of a found entity and distinguish it between other, similar entities</td>
</tr>
<tr>
<td>primary</td>
<td>access</td>
<td>selection</td>
<td>select, choice, assist in the choice, narrowing, limiting</td>
<td>to determine the suitability of entities and chose specific entities among them</td>
</tr>
<tr>
<td>primary</td>
<td>access</td>
<td>obtaining</td>
<td>acquire, obtain, delivery</td>
<td>to access the content of an information resource</td>
</tr>
<tr>
<td>primary</td>
<td>discovery</td>
<td>navigation</td>
<td>navigate, discover, explore, browse</td>
<td>to explore related entities, resources, domains, and the relationships among them</td>
</tr>
<tr>
<td>primary</td>
<td>discovery</td>
<td>recommendation</td>
<td>recommend, suggest, connect, reading advisory</td>
<td>to relate and suggest entities based on a particular shared attribute</td>
</tr>
<tr>
<td>primary</td>
<td>discovery</td>
<td>linking</td>
<td>linking data, linked data, linking information</td>
<td>to link to additional entity information, external to a particular metadata storage and retrieval system</td>
</tr>
<tr>
<td>primary</td>
<td>discovery</td>
<td>visualization</td>
<td>visualize, display, cluster</td>
<td>to visualize collections, entities or subsets based on certain attributes</td>
</tr>
<tr>
<td>auxiliary</td>
<td>education</td>
<td>education</td>
<td>educate, instruct</td>
<td>to educate users on how to use the catalog, how to interact with the bibliographic data, and educate users about topical domains</td>
</tr>
<tr>
<td>auxiliary</td>
<td>management</td>
<td>collection management</td>
<td>inventory, control, collection development, collection assessment</td>
<td>to list or register property, to assess collections, identify gaps, acquire, purchase, share, and plan the resources of a particular collection</td>
</tr>
<tr>
<td>auxiliary</td>
<td>management</td>
<td>sharing</td>
<td>share, exchange, interoperability, consistency</td>
<td>to provide consistency of data and facilitate the exchange of data or resources</td>
</tr>
<tr>
<td>auxiliary</td>
<td>management</td>
<td>tool assessment</td>
<td>discovery tool assessment</td>
<td>to facilitate the assessment of discovery tools and inform interface design</td>
</tr>
<tr>
<td>auxiliary</td>
<td>management</td>
<td>advocating</td>
<td>marketing, expression</td>
<td>to market library collections, express the value, reliability and authority of libraries</td>
</tr>
<tr>
<td>auxiliary</td>
<td>management</td>
<td>preservation</td>
<td>preserve, archive</td>
<td>to facilitate the long-term preservation and future usage of information</td>
</tr>
<tr>
<td>auxiliary</td>
<td>support</td>
<td>research assessment</td>
<td>institutional research assessment, reputation management</td>
<td>to provide measures and supporting evidence for research assessment of an institution or individuals within</td>
</tr>
</tbody>
</table>
The primary functions of bibliographic data fall under the categories of access and discovery of information. These are the main purposes bibliographic data and catalogs are designed to support. Among those are the finding, collocating, and choice objectives originally identified by Cutter (1876). More often we see that collocation is discussed as part of the finding function. Svenonius (2000) asserts that “the traditional finding objective specifies that what is to be found in a particular known document, while the traditional collocating objective specifies that what is to be found is a set of documents, defined by criteria such as author, work, and subject”. The IFLA FRBR model and later the LRM model integrate the two traditional concepts of finding and collocating into the one find user task (IFLA 1998, 2017). IFLA LRM defines find as “to bring together information about one or more resources of interest by searching on any relevant criteria” (IFLA 2017, p. 15). In this paper, our definition of “find” includes both finding a single or a set of entities and resources that meet certain criteria and therefore incorporate the collocation function in the finding function.

The identification function is described in the literature using definitions ranging from the traditional object which is to confirm an entity to be what it says it is based on certain criteria (IFLA 1998, Hider, 2017), to the expanded definitions of distinguishing one entity from another similar one (IFLA 2011; Svenonius 1992), and to understand the nature of resources found (IFLA 2017). Users need to first perform the finding function in order for the identification function to take place. According to Liu (2007) “metadata provides information that allows the user to identify appropriate resources without exploring the resources themselves. Metadata allows the user to evaluate an information resource's relative value or lack thereof” (p.12), which implies an evaluation component for the identification function.

Cutter’s “assist in the choice” maps to the selection function, a term used by the majority of the literature on the topic. Here again we see that the function of selection is only possible after the finding and identification functions are performed. This is also discussed by Harej and Zumer (2013) who indicate that "selection can be done only

<table>
<thead>
<tr>
<th>Auxiliary support</th>
<th>Curriculum support</th>
<th>Instructional assessment, curriculum support, instructional use, instructional integration</th>
<th>To support instructional assessment, link to education standards, and support curriculum and accreditation efforts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social interaction</td>
<td>Social interaction</td>
<td>Social connection, social interaction</td>
<td>To connect with other users, to enable communication among users about resources, to enable community building</td>
</tr>
<tr>
<td>Curation</td>
<td>Curation</td>
<td>Personal curation, institutional curation</td>
<td>To curate personal or institutional data for citation managers, reading lists, or other purposes</td>
</tr>
</tbody>
</table>
after identification: proper selection can be done only after we have gained sufficient information about the entities on which we will be performing selection” (p.752-753). The selection function allows users to only choose the entities or resources that are most relevant, accessible, or usable based on factors like availability, format, and equipment required to access content. Svenonius (1992) states, “[t]oday the choice or selection of a bibliographic entity is conditioned by a more complicated set of factors. For instance, an important attribute of non-book materials is the equipment needed to experience the materials. What computer system does one need to process a computer data file? What viewing equipment does one need to use a film?” (p. 12).

The obtaining function may often be the ultimate access goal of users, who are looking to acquire or gain access to the content of an entity or resource. This function is more easily facilitated by systems that link together the bibliographic data and the resources themselves, especially in a digital content collection (full-text, digital image, digital music content). In more traditional cases of library catalogs, when the bibliographic data are not directly linked to the content of the resource represented, the data need to offer adequate information for the user to gain access through physical means, request for loan, or purchase.

The navigation function has not always been discussed in the earlier literature as one of the functions of a catalog, although browsing or exploration of bibliographic data was possible even in a book or card catalog. Harej and Zumer (2013) differentiate finding from navigation by asserting that “[w]hile find corresponds to searching, explore represents browsing, which users often use to locate resources when they want to avoid formulating a search statement” (p. 744).

The access functions of bibliographic data that have not been extensively discussed, as such, in the literature include the recommendation, visualization, and linking functions. Relationships between entities, shared characteristics or a combination of them with other data (such as viewing or circulation) function as recommendations for similar entities and can facilitate ratings, rankings, and reader’s advisory services (Dali 2015; Dempsey 2012). Bibliographic data, such as class numbers of subject terms, can be used to visually represent (visualization) a collection or subset of a collection or domain. This function can also support the navigation function. In more recent years, we see extensive discussion of linking internal or local bibliographic data to external data that provide additional information about entities or facilitate navigation outside a particular system. As more linked data or data enhancement efforts make connections to external information, the linking function of the bibliographic data will become more prevalent.

One notable management function of the catalog discussed early on is its use as an inventory for collections. This inventory is used for collection management, including its assessment, development, aiding the acquisition of new resources, or removal.
of existing resources. Other auxiliary uses of bibliographic data that have been identified in the literature include functions such as educating the user of the catalog in the use of the bibliographic data included in them, sharing bibliographic data or resources, and preservation of resources and their content to ensure future access and use. Other uses of bibliographic data support external purposes. For example, the bibliographic data can be used to assess information retrieval systems, search engines, or other technologies (tool assessment), to provide evidence in support of the marketing of libraries, their value in society, and advocacy efforts (advocating). They are used to support assessment and curation needs of institutions or individuals, including research-related or curriculum-related activities (research assessment, curation, curriculum support). In addition, data like tags, reviews, or ratings serve as a means to facilitate user social interaction.

Conclusion

Our inventory of the complementary uses of metadata lay bare the multiple uses beyond Cutter’s objects. When we consider the value of metadata we are often concerned solely with user interaction. However, as we can see, there are many other ways in which metadata supports knowledge organization work – professionals’ work. A full understanding of these purposes and functions allows us to broaden our conceptualization of the work we do, and therefore, broaden the conversation about the value of the work of knowledge organization.

References


**Additional sources used in the analysis**


José Augusto Chaves Guimarães, Maria Claudia Cabrini Gracio, Daniel Martínez-Ávila, Rodrigo de Sales

The Spirit of inquiry’s power to influence in 21st-century KO research: Jesse Shera and Margaret Egan

Abstract
This study aims to examine the influence of Jesse Egan and Margaret Shera on scientific production in knowledge organization (KO) based on the citations that these authors have received in the literature. Thus, we conducted an ego-centric citation analysis of both authors in co-authorship as their collaborative effort was responsible for some of the most fruitful theoretical foundations for library and information science and knowledge organization. In this vein, we searched records that cited works co-authored by Shera and Egan in the database Scopus and, after making a selection of the citing papers whose titles suggested a link to knowledge organization, we analyzed the authors that were most co-cited with Shera and Egan, the main venues for publication, and the years of publication in order to analyze the permeability (power to influence) of the two authors in the KO field. The results reveal not only the effective influence of Egan and Shera on the theoretical-methodological universe of KO but also and more specifically on the ISKO universe.

Introduction
The Graduate Library School of the University of Chicago (GLS-UC) (1926-1989) played a special role in the scientific development of the Library and Information Science (LIS) field. This school institutionalized a "spirit of inquiry" (Richardson Jr. 1982) based on the German tradition of seminars (in opposition to the French tradition of lectures) as dialogical spaces for methodical inquiry in the search for scientific truths that are independent of their utilitarian applications (Richardson Jr. 2010). This aspect contributed to the construction of a kind of episteme of a scientific field which Pierce Butler (1933) called "library science,” the scientific study of books and users following a social-scientific approach, directly influenced by the ideas of John Dewey (1929) concerning identification of the sources of a science of education. This conception of library science also brought new techniques from the quantitative social sciences to the research problems of librarianship.

In this context and representing a second generation of GLS-UC faculty, Jesse Shera and Margaret Egan established an important scientific collaboration, especially in aspects related to the field of knowledge organization (KO). Egan and Shera (1952) coined the term “social epistemology” as a field of study related to the social production, distribution, and utilization of intellectual products in such a way that “informed social action” acts as the main goal of library service, by means of the evaluation of bibliographic services; and supported by a theoretical framework from the study of information-seeking behavior, knowledge organization, and bibliometrics (Furner 2004; Zandonade 2004). Shera and Egan (1956) also brought other
contributions to the North American Knowledge Organization literature such as the
general principles for the construction and maintenance of systematic or classified
catalogs, an “uncommon device of American bibliothecal organizations” (Custer 1957,
p. 122), that were more devoted to the Cutter’s tradition of dictionary catalogs.

Our paper aims to analyze the influence of Egan and Shera on scientific production
in Knowledge Organization based on the citations that these authors receive in the
literature.

Methodology

Our methodology is based on the idea of “the power to influence” used by Dietmar
Wolfram (2016) in knowledge organization for his ego-centric citation analysis of Hope
Olson (and whose most well-known work “The power to name” inspired Wolfram’s
title). Ego-centric citation analysis (White 2000) was used as a way to verify “the
influence of an individual, publication or journal based on the references they provide
in their own work, the citations their works receive and the works and people with
whom they are co-cited” (Wolfram 2016, p. 332), following three approaches: citation
identity (all the authors that a citing author cites), citation image makers (authors who
cite a certain author), and citation image (authors with whom a certain author has been
cocited). Other examples of ego-centered bibliometric analyses of prominent
knowledge organization authors include those of MP Satija (Swain 2009), S.R.
Ranganathan (Smiraglia 2013), and Paul Otlet (Rodríguez-Bravo et al. 2017). While
citation analysis can be considered a way to represent social knowledge
organization systems (Hjørland 2013), we believe our paper also falls under the category of domain
analysis, bibliometric studies being one of the valid approaches to domain analysis
sanctioned by Hjørland (2002, 2017). For the purpose of this paper, we consider Shera
and Egan together as an individual for analysis (an ego), especially in relation to the
relevance of their scientific collaboration and not only Jesse Shera for the development
of the KO field (Furner 2004).

Accordingly, we searched records that cited works co-authored by Jesse Shera and
Margaret Egan in the database Scopus using the following retrieval profile:
[REFAUTH (shera) AND REFAUTH (egan)]. By October 17th, 2017, we had retrieved
13 records co-authored by Shera and Egan, which after disambiguation resulted in five
works receiving a total of 82 citations: “Foundations of a theory of bibliography”
(1952), with 56 citations; “A review of the present state of librarianship and
documentation” (editions of 1953, 1961 – in Portuguese –, 1966, and 1971), with nine
citations (seven citations for the 1953 edition and one citation for the other editions);
“Prolegomena to bibliographic control” (1949) and “Bibliographic organization”
(1952), with six citations each; and “The classified catalog; basic principles and
practices” (1956), with 5 citations. In a second stage, we selected the citing papers
whose titles suggested a link to knowledge organization and analyzed the authors that
were most co-cited with Shera and Egan, the main venues for publication, and the years of publication in order to analyze the permeability (power to influence) of the two authors in the KO field. In other words, we considered this four-year period of scientific collaboration between Egan and Shera as an important domain to be analyzed and a seminial space for KO studies.

**Results**

Out of the 81 citations to the five works, 35 (43%) appear in the area of Knowledge Organization. The majority of them (80%) are in documents published after 1996, and 58% of them published in documents from 2001 to 2010. These citations suggest a link between the work of Shera and Egan and the current research in Knowledge Organization.

Figure 1 presents the diachronic analysis of these citations showing the preference of The Classified Catalog and Prolegomena to Bibliographic Control until 1985; a balance from 1986 to 2005, and the predominance of Foundations of a Theory of Bibliography and Bibliographic Organization after 2011.

![Figure 1: Distribution of citations to Shera and Egan’s works](image)

Out of the 55 citations to "Foundations of a Theory of Bibliography" 14 citations (25%) appeared in those papers whose titles suggest a connection with knowledge organization. Among the citing authors, Jack Andersen and Jonathan Furner are the contributors with the highest number of citations, with five citations each (and together accounting for 18% of the citations). As for the sources, the majority of these citations (36%) appeared in ISKO publications - the journal "Knowledge Organization" and the "Advances of Knowledge Organization" – (36%). The presence of 72% of the citations in ISKO publications is especially significant considering the relevance of ISKO as an
international scientific space for the KO field. It is important to stress that this high number of citations also reveals an effective concern of the ISKO environment with the historical foundations of the field.

The most co-cited authors with Egan and Shera in these papers were Birger Hjørland (16 citations), Jack Andersen and Jenna Hartel (nine citations each), and John Budd and Michel Foucault (eight citations each). It is interesting to observe that all of the most co-cited authors mentioned above are mostly known for their theoretical studies. This suggests that the impact of Egan and Shera was more theoretical than practical, something that is consistent with the investigative and reflective proposal of the GLS-UC.

Out of the nine citations to "A review of the present state of librarianship and documentation" five citations (56%) appeared in papers that we considered to be on knowledge organization and single-authored by the following people: J. Furner, T. Zandonade, C.K. Malone, J. Warner, and N. Oddone, N. These citations appeared in the following journals: “Library Trends” (two citations), “IEEE Annals of the History of Computing” (one citation), “Journal of Librarianship and Information Science” (one citation) and “Ciência da Informação” (one citation). The citation in the IEEE Annals of the History of Computing reveals the relevance of the work of Egan and Shera not only for LIS, but also the recognition of this relevance from the history of Computer Science. The main authors co-authored with Shera and Egan in these papers were: S. Fuller (12 citations), W.B. Rayward, and Sambaquy L. de, Q. (six citations each), and J. Farradane, T.J. Froehlich, W. Goffman, and R.R. Shaw (four citations each).

All 6 citations to "Bibliographic Organization" appeared in papers identified as knowledge organization and mostly (83%) published during the period 2004–2012. The citing authors in these publications are: E. Milonas (two citations) with one citation in two papers, and Y. González Pérez, E.K. Jacob, T. Lincoln, and T. Zandonade, with one citation in one paper. The sources in which these citations appear are: "Library Trends" (two citations), "Acimed" (one citation), "Advances in Classification Research Online" (one citation), "Advances in Knowledge Organization" (one citation) and Cataloging & Classification Quarterly" (one citation). The main co-cited authors with Egan and Shera in these papers are: S. Fuller (12 citations), E.K. Jacob (six citations), T.J. Froehlich, W. Goffman, and B. Kules (four citations each), and V, Broughton, F. Machlup, and W.B. Rayward, E. Rosch, and B. Shneiderman, (three citations each). The wide range of citing and co-cited authors as well as of citing journals reveal the different ways in which the work of Egan and Shera has been relevant within the knowledge organization literature.

Out of the six citations to "Prolegomena to bibliographic control", five citations (83%) appeared in papers considered to be on KO according to the titles. M.L. Mackenzie authored two of these papers, and J. Furner, H.H. Wellisch, and T. Zandonade each authored a paper citing this work. These papers were published in the
following titles: "Library Trends" (two citations), "Journal of the American Society for Information Science" (one citation), "Library and Information Science Research" (one citation), and "Proceedings of the ASIS Annual Meeting" (one citation), during the period 1980-2004. The most co-cited authors include: S. Fuller (with 12 citations), W. Goffman, and L.B. Heilprin (with four citations each) and M. Kaltenbach, F. Machlup, and W.B. Rayward (with three citations each).

All 5 citations to the work "The classified catalog" appeared in papers in knowledge organization and single-authored by the following people: L. Butler, T.H. Connell, R.L. Liske, N. Roberts, K. Rondestvedt, and L. Summerville. These citations appeared during the period 1972-1998 in the following sources: "Cataloging & Classification Quarterly" (two citations), "Audio Visual Communication Review" (one citation), "Journal of Internet Cataloging" (one citation), and "Social Science Information Studies" (one citation). The most co-cited authors with Shera and Egan in these papers were C. Palowitch, with 3 citations, and G. Barhydt, M.J. Bates, E. Coates, A.C. Foskett, D. Haykin, F. Riggs, C. Schultz, V. Sessions, and M. Taube, with two citations each. These results might be unexpected if this is considered the most specific work by Egan and Shera in the KO field. On the other hand, the low number of citations confirms that the influence of Shera and Egan was greater on the theoretical dimension of KO than on the practical side, as this book was devoted to the steps and procedures for developing a classified catalogue.

Figure 2 shows the co-citation network of Shera & Egan's works and the most co-cited authors with these works. We observe that five authors (in green circles) are co-cited with three of Shera & Egan's works, thus being the main contributors to the creation of the Shera & Egan's citation image. Among them, we highlight Steve Fuller with the highest intensity of co-citation (shown in the figure by the thickness of the connection with the works). It is also worth mentioning the intense co-citation of Hjørland with the work “Foundations of a Theory of Bibliography”. As for the other works, with the exception of Classified Catalog, all of them were co-cited with a wide range of co-authors.
Results & conclusion

The results show that, among all the authors citing these five works, Tarcisio Zandonade is one of the most important contributors (image creator) to the image of Shera and Egan as scientific collaborators, as he cited three of the analyzed works by them: "A Review of the Present State of Librarianship and Documentation", "Bibliographic Organization", and "Prolegomena to Bibliographic Control". Jonathan Furner also plays an important role as he used two of the analyzed works in his research: "A Review of the Present State of Librarianship and Documentation" and "Prolegomena to Bibliographic Control". Finally, Jack Andersen is also an important image creator as he based several of his papers on "Foundations of a Theory of Bibliography".

Citations appeared more frequently in the ISKO publications – "Knowledge Organization" and "Advances of Knowledge Organization" –, the journal "Library Trends", and, to a lesser extent, the journal “Cataloging & Classification Quarterly”. The authors that were most co-cited with the analyzed works by Shera and Egan are: Steve Fuller, Birger Hjørland, and William Goffman. Other important authors that also contributed to the creation of the citation image of the scientific collaboration between Shera and Egan, although with less co-citations, include: Jenna Hartel, W. Boyd Rayward, Jack Andersen, John Budd, Michel Foucault, and Thomas J. Froehlich.

In general, the effective influence of the Egan and Shera collaboration on the theoretical-methodological universe of KO is patent, especially considering the universe in which the works are most cited (the ISKO publications) and the
epistemological nature of the authors that cite them. This aspect confirms the pioneering character of the GLS in terms of academic conception and research orientation, which, even some decades later, would still find a fertile environment for reflection in the ISKO.

References
Hjørland, Birger (2013). Citation analysis: a social and dynamic approach to Knowledge Organization. *Information Processing and Management* 49: 1.313-1.325.
Shera, Jesse Hauk and Margaret E. Egan (1956). *The Classified catalog: basic principles and practices*. Chicago; ALA.


Jonathan Furner

Truth, relevance, and justice: towards a veritistic turn for KO

Abstract
Analysis of recent work at the intersection of epistemology and ethics points to a potentially innovative mode of critical knowledge organization (KO) that (1) is informed by applied social epistemology, (2) is inspired by values of epistemic justice, (3) is respectful of human rights, and (4) privileges truth over relevance.

1. Introduction
In efforts to construct theoretical foundations for information studies, scholars have drawn variously on conceptions of social epistemology, social justice, and epistemic justice (among other ideas). Is it possible to untangle the relationships among these conceptions, in order to arrive at a compelling justification for a distinctively critical LIS (library and information science, generally) and KO (knowledge organization, specifically)? In this paper, I address this question through a conceptual analysis of truth, relevance, justice, and related notions, and point towards the possibility of a veritistic turn for KO.

2. Ontology
One way to start thinking about KO is to consider it as a field of inquiry, and to attempt to define it by identifying a primary goal shared by its practitioners. From an LIS perspective, for example, we might decide that the goal of KO is to establish a set of principles by which tools and techniques for the access-oriented description of resources – tools and techniques such as controlled vocabularies, metadata schemata, and ontologies – may be designed, constructed, and evaluated. From a standpoint with a wider view, contrastingly, we might choose to relax the pragmatic requirement that KO should be directed towards helping us build information services, and at the same time to recognize that KO may usefully be defined in comparison with the philosophical field of ontology.

Ontology is commonly characterized as the philosophy of being – of what there is, or what kinds of things there are. Along these lines, the practical work of ontology may be construed as the determination of the ways in which individual things are related to one another, and thus are organized into kinds or classes. Using the same template, but replacing things with the narrower category data, produces a formulation that looks as if it could be put to work as a definition of KO. If we understand our resource descriptions and authority records, our catalogs and data models, our indexes and thesauri, our classification schemes and terminological databases, as themselves being made up of items of data – i.e., expressions of propositional knowledge – then it is
surely not too much of a stretch to consider the practical work of KO as the
determination of the ways in which items of data are related to one another, and thus
are organized into kinds or classes. Similarly, here lies an opportunity to explore the
potential equivalence of KO and philosophy of data (see, for example, Furner 2017).

3. Epistemology

It might be argued the subject matter of philosophy of data is primarily comprised
of elements of three more-traditional branches of philosophy, themselves bounded
more-or-less arbitrarily and with considerable overlap, that we might conveniently label
philosophy of mind, philosophy of language, and philosophy of belief. I use the term
epistemology, traditionally employed to refer to philosophy of knowledge, to denote
this last, slightly broader category, so that I can make a distinction between two generic
families of theories of belief.

One approach to explaining the difference between these families involves first
making an ontological commitment to at least two categories of abstract objects,
namely attitudes and propositions, and then conceiving of beliefs simply as attitudes
towards propositions. Once this is done, truth-oriented theories may be defined as
theories of belief that distinguish between true and false beliefs, that is, between beliefs
that true propositions are true, and beliefs that false propositions are true, respectively;
while relevance-oriented theories may be defined as theories of belief that distinguish
between relevant and non-relevant beliefs, that is, between beliefs that relevant
propositions are relevant, and beliefs that non-relevant propositions are relevant.

We may use this distinction to express a historical disconnect between epistemology
as a subfield of analytic philosophy, and other fields such as library and information
science. The theorizing about belief that occurs in epistemology is typically truth-
oriented; that done by LIS scholars has typically been relevance-oriented. Specifically,
the significance of perceived relevance as a criterion for determining the retrieval-
worthiness of resources comprised of aggregations of expressions of propositions is
well-documented (see, for example, Borlund 2003).

4. Social Epistemology

Several other fundamental distinctions are helpful for characterizing the various
branches of the field of epistemology. We may distinguish between pure epistemology
and applied epistemology, even if those labels are seldom used in practice – i.e.,
between theories that are “merely” descriptive of the nature of doxastic and epistemic
concepts and practices, and those that are normative in the sense that they seek to
specify the practices that are most conducive to believing true or relevant propositions.

Regardless of whether one’s objectives are descriptive or normative, one’s approach
or methodology may be more-or-less rationalistic, and more-or-less naturalistic,
depending on one’s readiness to admit different kinds of evidence in support of one’s
conclusions. Similarly, one’s interests may be more-or-less individualist, and more-or-less social, to the extent that one chooses to focus on interpersonal interaction as a factor in the formation of beliefs.

Applied social epistemology, then, is the study of normative questions about the social practices that are most likely to generate true or relevant beliefs (see, for example, Goldman 1999). So, for example, we might ask, On what kinds of grounds should we assign positive evaluations to testimony? Under what kinds of conditions should we believe that what we read is true, or that what we’re told is relevant?

5. Justice

Like truth and relevance, and like other concepts such as beauty and freedom, justice is a value – a more-or-less desirable feature of the outcomes of people’s decisions and actions. The nature of justice has been the subject matter of a huge body of philosophical literature spanning ethics, social and political philosophy, and jurisprudence, with both descriptive and normative dimensions (see, for example, Sandel 2009). Simplifying greatly, one theme that has attracted much attention is the idea of justice as fairness, with its corollary that the kinds of practices that generate states of justice most effectively are those in which people are treated fairly.

Of course, simply substituting “fairness” for “justice” does nothing to explain the concept or to suggest what kinds of criteria can be used to evaluate how fair or just any given treatment might be. Yet we may recognize that, typically, justice is seen to be done when people are treated in accordance with their just deserts, on their merits or needs, without prejudice or bias or discrimination, without violation of their human rights, without limitation of their freedoms, and without the exercise of any form of oppression stemming from asymmetric power relations.

Concepts of justice are commonly incorporated into rights-based ethical frameworks through the identification of natural, human, civil, group, or individual rights to equitable access to goods or opportunities, and to fair treatment by others, free of impediment and obstruction. Such rights include those that are attributed to speakers and hearers, knowers and inquirers, teachers and learners – in other words, to all of us, as communicative, social beings: (1) rights to think, to conceptualize, to categorize and classify, to believe and have opinions; (2) rights to express, to give voice to one’s thoughts in speech, writing, and other forms; (3) rights to access, seek, inquire, find, hear, and know the thought and expression of others; (4) rights to be heard, to publish and broadcast, to reach an audience without being censored, silenced, hidden, or ignored; (5) rights to be left alone, to maintain one’s privacy, to be forgotten; (6) rights to be believed, to be given credence, to be treated as credible; inter alia.
6. Social Justice

Different theories of justice account for the relationships between notions of desert, merit, bias, and so on, in different ways. Theories of distributive justice focus on the outcomes of actions taken to distribute quantities of resources among the members of given populations. Theories of social justice, which may or may not simultaneously be theories of distributive justice, highlight the importance of individuals’ identifying with certain groups – races, genders, and classes, for example – and of ensuring that such memberships are taken into appropriate account in any calculus of justice (see, for example, Roberts and Noble 2016).

Different theories of social justice work with different ideas about how “appropriateness” in accounting can be determined. Many of us in KO education avow a commitment to social justice in distributive terms, which means that in general our goal is to change the world in ways that include (1) reducing divides, disparities, and inequalities between rich and poor, and between powerful and powerlessness, (2) generating fairer distributions of social, cultural, economic, and political opportunities, and (3) building and maintaining thriving communities in which basic human rights and freedoms are respected. Translating that specifically into a vision of the future of information services, we say that we are working towards a society (1) where members of all social groups enjoy equitable access to the knowledge they need to get their jobs done and live their lives to the full, (2) where everyone, not just an elite minority, is empowered to preserve their own artifacts and stories and ideas, in the ways they wish, and to the extent they wish, and (3) where the actions of governments and corporations respect the rights of individuals and communities to intellectual freedom, to privacy, and to cultural property.

Yet, to focus on distributive social justice alone would be to ignore one of the most prevalent kinds of injustice in our society, which is that caused by the systemic oppression of, or discrimination against, specific social groups. Iris Marion Young (1990) distinguishes between what she call the “five faces of oppression”, pointing out that injustice is manifested as exploitation, whenever labor is divided in such a way that women work specifically in order to maintain the power, wealth, and status of men; as marginalization whenever those who are old, or young, or poor, or disabled, or otherwise dependent on others, are thereby deprived of basic rights and freedoms; as powerlessness whenever working-class people find that they are unable to participate in making the decisions that have the greatest effect on their lives; as cultural imperialism when the very means of interpretation, expression, and communication in our society are so completely controlled by white Christian heterosexual men that the experiences and values of nondominant groups are rendered invisible; and as violence when we choose to tolerate the fact of black people living their lives under the constant threat of harassment, intimidation, and physical violence simply on account of their
group identity.

Working towards social justice as a goal therefore involves the basic reform of oppressive, discriminatory social practices and institutions, as well as the redistribution of resources. Among those practices and institutions are those by whose means we produce and consume knowledge—the practices and institutions, in other words, of library and information services, including knowledge organization systems (KOSs) such as bibliographic classification schemes, subject heading lists, and thesauri.

7. Epistemic Justice

Among theories of justice, theories of epistemic justice are special in the way that they focus on the fairness of our treatments of people in their capacity as believers and as knowers (see, for example, Fricker 2007). These theories, too, may or may not be theories of distributive justice, depending on whether or not they suggest how quantities of what we might call epistemic resources, such as data or knowledge, or opportunities to access such resources, may be distributed fairly. Similarly, they may or may not be theories of social justice, depending on whether or not they emphasize people’s affinities with social groups as factors to be weighed when determining the fairness of particular treatments.

Along these lines, Miranda Fricker (2013, p. 1.318) distinguishes between two kinds of epistemic injustice. Distributive injustice occurs whenever epistemic resources, “goods such as education or information,” are distributed unfairly. Discriminatory injustice (which, had she seen reason to follow Young, Fricker might instead have called oppressive injustice) occurs whenever wrong is done to an individual either as a testifier or as a sensemaker—that is, either as a potential source of evidence, or as a potential acquirer of true belief: testimonial injustice happens “when a speaker receives a deficit of credibility owing to the operation of prejudice in the hearer’s judgment” (and Fricker gives the example of a police officer not believing the testimony of a young black male; p. 1.319), whereas hermeneutical injustice happens “when a subject who is hermeneutically marginalized (that is, they belong to a group which does not have access to equal participation in the generation of social meanings) is thereby put at an unfair disadvantage when it comes to making sense of a significant area of their social experience” (and Fricker gives the example of a female victim of domestic violence unable to process the reality of their situation; p. 1.319).

Whereas the goal of social justice has, over the last decade or so, become quite a commonly articulated objective for the providers of library and information services—at least at the more activist or progressive end of the spectrum of views about the proper role of LIS institutions such as KO schemes—the notion of epistemic justice has not been taken up by LIS scholars or practitioners to anything like the same degree. This is curious, given the magnitude of the overlap that may be discerned in the respective
concerns of KO and epistemology, especially applied social epistemology.

8. A Critical KO

Analysis of recent work at the intersection of epistemology and ethics, such as that highlighted in the previous section, points to a potentially innovative mode of critical KO, in four related respects.

(1) A KO that is informed by applied social epistemology

In the first place, we may clarify how the goals of KO may be viewed as the result of applying the normative theory of a particular flavor of social epistemology. Answers to questions of the kinds addressed by applied social epistemology – such as those directed to identifying the conditions under which testimony should be evaluated as true or relevant – provide sets of desiderata for the design both of our search engines and of the KO systems that underlie them.

(2) A KO that is inspired by values of epistemic justice

Secondly, we have an opportunity to characterize our values with more precision. Notwithstanding the virtue inherent in its pursuit, to present social justice as the primary end to which library and information service is directed is to undermine the unique character of such service. That unique character is captured in the idea that the librarian’s mission is to provide access to the world’s recorded knowledge: in other words, to foster epistemic justice by enabling the dissemination and acquisition of true beliefs.

(3) A KO that is respectful of human rights

Thirdly, we have an opportunity to articulate a right to be believed – a right to testimonial justice – to set alongside the rights to free thought and free expression that are already encapsulated in documents such as the UN’s Universal Declaration of Human Rights (United Nations, 1948) and in IFLA’s Code of Ethics for Librarians and Information Workers (International Federation of Library Associations and Institutions 2012). Recognition of a need to respect the rights to testimonial justice held by members of a marginalized, oppressed, or previously underrepresented community could be the catalyst for long-overdue revision of KO systems whose developers have not attributed appropriate levels of credibility to testifiers from that community.

(4) A KO that privileges truth over relevance

Fourthly, we have an opportunity to proclaim a veritistic turn, in the course of which the centrality of epistemological concerns to LIS and KO is recognized, and truth supplants relevance as a core value. It seems both possible and desirable to distinguish between relevance-oriented and truth-oriented characterizations of the mission of the information worker, along the same lines on which it is useful to distinguish between two families of theories of belief. Relevance-oriented KO is that which seeks to evaluate its practices, institutions, and products on the basis of the extent to which the
desires of users are satisfied; truth-oriented KO is evaluated on the basis of the extent to which the beliefs acquired by users are true.

9. Conclusion

Given the historical attachment of LIS to relevance-oriented service, the influence of postmodernist denials of the possibility of objective knowledge, and the maintenance in IFLA’s Code (among many others) of a statement of information workers’ commitment to “neutrality,” this last conclusion may be the most controversial of the three. The skeptic might even wonder if the controversy is simply a contemporary manifestation of the age-old debate in librarianship over the competing principles of “Give them what they say they want” and “Give them what we think they need”. I suggest that the most critical task facing KO theorists today is to recognize the moral emptiness of both of those positions and provide a justification for a KO that is consistent with contemporary, pluralist conceptions of truth.

References


Verbal protocols in Brazilian Information Science: a perspective from indexing studies

Abstract
Verbal protocols have unlimited applicability as an introspective technique for collecting qualitative data from scientific research in information science. Protocol analysis of verbal thought reports is frequently used in different knowledge fields, essentially as it enables observation of the phases in subjects’ cognitive processing through verbalizations made during reading activities. As a way of increasing knowledge about the verbal protocol as a methodological resource in information science, different modalities of the technique were applied to Brazilian research. It was verified that research conducted by verbal protocol is, for the most part, aimed at representing thematic information, focusing particularly on inquiries that involve document reading and professional practice of indexing processes and subject cataloging in a university library. Although such research areas are the focus of the present study, the introspective nature of the verbal protocol method favors further research into other issues, especially in the area of knowledge organization and representation.

1. Introduction
The cognitive revolution of the 1960s stimulated scientists’ interest in knowledge about how thinking allows subjects to generate solutions to perform new procedures, that is, apply previously acquired knowledge to solve new problems. This context of enthusiastic technological innovation contributed to the resumption of introspection in psychology, until then discredited as a scientific method. Introspection techniques are classified into three groups: self-observation, in which research subjects report their own mental events – that is, introspection whereby subjects observe and analyze their behavior; self-report or self-perception, in which subjects verbalize their experience to the researcher – that is, retrospection; and “High Thinking”, a method put forward whereby research subjects think aloud while performing a certain activity – a technique known as verbal protocol or protocol analysis (Radford and Burton 1974).

Based on the introspection method, Anders Ericsson and Herbert Alexander Simon, researchers in the field of experimental cognitive psychology, indicated in the 1980s that subjects’ cognitive processes can be described as sequences of served and structured information. To do so, they recommend the use of protocol analysis of verbal reports of thought sequence, also known as records of verbal reports or verbal protocols, whereby observations of the cognitive process provide information about the phases of individual processing, that is, the successive states of information verbalized by the subject while accomplishing a given task.

Used as an introspective method of data collection, verbal protocols have been shown to be a valuable means of observing conscious mental processes, and are the
main data source for cognitive research, being used in different areas of knowledge since they enable observation of the subject’s cognitive processing phases by means of thought verbalizations. A growing amount of scientific research into verbal protocols has been conducted since 1975 in areas such as psychology, education, medicine, computer science, ergonomics, information science and librarianship, nursing, and gerontology, among others (Alonso-Arroyo, Fujita, Gil Leiva and Pandiella 2016).

Observations of mental processes provide important information about individual processing steps while subjects exteriorize their mental processes during an activity and maintain the sequence of processed information. Therefore, “High Thinking” or the subject’s *Thinking Aloud* while performing a task is recorded and transcribed verbatim, producing verbal protocols, that is, verbal reports of the conscious mental processes of the research subject (Cavalcanti and Zanotto 1994).

As a methodological resource, verbal protocols can be applied in three modalities. Using the *Individual Verbal Protocol*, the researcher intervenes with a single objective, to remind the participant subject about the importance of “High Thinking” in the execution of an activity, without performing any other types of interference or comment. On the other hand, the *Interactive Verbal Protocol* involves the active participation of the experienced researcher who, by motivating the subject during the activity, contributes to the achievement of more satisfactory levels of understanding. This interactive and innovative verbal protocol format was proposed by Nardi (1999) when demonstrating the possibility of interaction in protocols within the area of applied linguistics. In turn, the *Group Verbal Protocol*, another modality derived from the above-mentioned author’s research, arose from adaptations in the technique to investigate the opinion of a certain group of subjects on a specific theme, focusing on socially constructed cognition. The conduct and effectiveness of verbal protocol modalities depends, to a large extent, on the researcher’s position at the time of their application. According to Nardi (1999), the type of researcher participation can be divided into: passive participation (does not interact with the other participants, being a simple observer); moderate participation (alternating between observer and active participant roles); active participation (seeks to do what other participants do); and full participation (one of the participants themselves who decides to analyze the group data).

The complexity that permeates professional practices in information science and librarianship demands special attention to cognitive strategies, and protocol analysis of verbal thought reports is a timely alternative for uncovering subjects’ conscious experience in information processing. In Brazil, this choice of technique as an option for data collection and pedagogical resource in undergraduate education and continuing education has a significant presence in research into the study and observation of document reading processes in indexing and indexers’ professional practice.

To broaden knowledge about verbal protocols as a research resource in information
science, bibliographical research was carried out to identify Brazilian research projects that chose to use verbal protocols to observe and analyze the activity of subject indexing. The objective of this study is to present an overview of the research into indexing being done using verbal protocols and of the advances made in Brazilian academic production involving studies undertaken with this approach.

2. Research into subject indexing with verbal protocols

In the context of Brazilian information science, verbal protocols are most frequently applied in research aimed at observing reading processes. Research which applies the individual verbal protocol is directed to the observation of reading strategies used by indexers and subject catalogers during the representation process of document information. According to Fujita, Nardi and Fagundes (2003a), reading is a basic activity of document analysis, which is linked more directly to indexing objectives, that is, a professional reading that encompasses cognitive and metacognitive strategies, the use of the introspective technique of the verbal protocol thus being an effective way of finding out indexers’ individual processing. In the authors’ view, the verbal protocol “... provides direct access to the reading mental process while being carried out by the reader, unlike the ones that reveal only reflection after the reading process” (Fujita, Nardi and Fagundes, 2003b).

Accordingly, Fujita (2009) carried out a study which analyzed a socio-cognitive context to verify the perception of subject catalogers and university library users regarding book indexing, exposing the distance that exists between the literature that has been produced and the professional practice of librarians in this environment. In another instance, the idea was proposed to investigate the document reading of subject catalogers for book indexing, whereby the technique was applied to subjects from a sample of nine university libraries, leading to the document reading model being adapted for book indexing as a quality tool for representing concepts during subject analysis (Fujita 2013).

As a way of contributing to research on document languages, Boccato (2009) applied verbal protocols among catalogers, librarians and first and final year students of Civil Engineering, Pedagogy and Dentistry courses at São Paulo State University, with the purpose of verifying the process followed in the activities of indexing and conducting a bibliographic search in the online collective catalog of the institution’s Library Network.

Sousa (2012) studied the thematic representation of books that takes place during the process of subject analysis in the cataloging of subject matter for production of document information in the library catalogs of the Federal Institutes of Education, Science and Technology through a study which observed cataloging librarians. Reis (2012) used the technique to understand and analyze catalogers’ document reading and thus analyze textual structure in the subject cataloging of books, by observing the
expertise they have acquired from the profession, within the specific domain of university libraries. To contribute to the professional functioning of classification as an interpretative process, focusing on document subjects and covering users’ needs within the academic community, Sousa and Fujita (2013) considered catalogers’ perceptions in order to investigate the thematic description of book documents from the perspective of the bibliographic classification process. According to Redigolo (2014), the technique revealed the catalogers’ cognition about the subject analysis of books in Brazilian and Spanish university libraries, which favored a comparative analysis between these two realities and the elaboration of guidelines to develop the subject analysis in this informational context.

Tartarotti (2014) used the method to collect data from librarians at the University of São Paulo working on information retrieval systems in specialized scientific areas of health for comparative observation of the cognitive processes of document analysis, of the subject thesis cataloging in online collective catalog Dedalus and the indexing process in the Lilacs database. In Silva and Boccato (2012), the application of modality with interaction and the researchers’ moderate intervention aimed to evaluate the use of university libraries’ online collective catalogs in the information retrieval interface and in the users’ socio-cognitive context. The study by Andrade (2014) is one of the few carried out in the archival environment, whose use of the modality made it possible to analyze users’ information search strategies from the Judiciary Section of the Federal Trial Court of Paraíba State (JFPB).

Regarding interactive verbal protocols, this modality has been applied more in the teaching of subjects such as indexing and document reading in undergraduate courses in librarianship. The results of this didactic option have been presented through some scientific papers, such as the study by Fujita, Redigolo and Dal'Evedove (2007) which set out to evaluate the use of this modality as a pedagogical resource in the teaching of indexing to indexer apprentices; Borba and Fujita (2005) and Borba (2006) analyzed the role of the experienced teacher in initial indexer training through the activity of document reading in the classroom with students studying librarianship at the Universidade Estadual Paulista, with a view to providing a methodological description of document reading in undergraduate courses and information systems; and more recently by Fujita (2009) who investigated, from the perspective of the interactive verbal protocol, the applicability of the document reading model for indexing scientific articles as a socio-cognitive teaching methodology for apprentices.

Group verbal protocols are applied fairly widely, and used frequently in areas such as document reading, indexing policy and document languages from a socio-cognitive perspective, as well as being a pedagogical resource in the classroom, an approach presented by Fujita and Rubi (2007).

Among the research conducted to date, there is Gonçalves (2008) who observed the socio-cognitive context of users who are members of research groups of the
Universidade Estadual Paulista. Librarian and users have also been the object of study, whose use of the modality served to observe the context of socio-cognitive indexing in university library book cataloging (Fujita 2009).

Boccato (2009) evaluated the use of document language in online collective university library catalogs by collecting data with leaders, reference librarians, cataloguing librarians, undergraduate students and leading teachers or research group members in order to obtain the participants’ social opinions on the performance of document languages in relation to the needs of thematic treatment and information retrieval. On the other hand, Martinho (2010) resorted to reading as a social event with theoreticians and professionals of the area to verify how the principles advocated by Cutter influenced the information representation processes, due to the imprecision in process definition and conceptualization of subject cataloging before the other processes of thematic information treatment.

In Dal'Evedove and Fujita (2012), the authors investigated the reality of the process of cataloging the subject in different university libraries from the perspective of the librarian, whose results led to further research into the subject cataloger’s professional knowledge regarding the role of indexing policy in the context of university libraries (Dal’Evedove and Fujita, 2013).

Dal'Evedove (2010) analyzed the daily actions of leading librarians, reference librarians and subject catalogers, based on the knowledge and process of thematic information treatment in the context of university libraries. In another situation, the group voice protocol helped in validating a set of index policy guidelines for university libraries (Dal’Evedove 2014). In addition, in the indexing policy area, the works of Rubi are highlighted (2004; 2008) which analyze the context of library indexing and professional knowledge about the indexing policy, resulting in an indexing policy proposal with philosophy and guidelines for accomplishing the indexing process in university library cataloging.

Fujita and Rubi (2006) used reading as a social event to obtain indexing policy subsidies through indexers’ organizational knowledge of university libraries; Rubi, Fujita and Boccato (2012) did likewise, with the aim of transforming the indexer’s tacit knowledge into explicit knowledge in the form of an indexing manual. In Kochani, Boccato and Rubi (2012) and Kochani (2014), the modality served to discuss and outline policy guidelines for clippings in automated systems of Social Communication Coordinators in Brazilian university environments.

3. Final considerations

The verbal protocol is a methodological resource widely used in Brazilian information science, especially to observe document reading and the professional practice of librarians in subject indexing activities. Indexing research using verbal protocols as a teaching and/or research technique presents consistent results.
It is observed that most of the research conducted with a verbal protocol is related to issues involving document reading and professional practice in the indexing and subject cataloging processes of a university library, since it enables observation of the phases in the subject’s cognitive processing through verbalizations made during reading activities. Indexing searches by applying individual verbal protocols are essentially based on observing reading strategies used by indexers and subject catalogers during the process of representing document information. Interactive verbal protocols are more developed in the teaching of subjects such as indexing and document reading in undergraduate courses in librarianship. The modality of group verbal protocol is used frequently in areas such as document reading, indexing policy and documenting languages from a socio-cognitive perspective, as well as a pedagogical resource in the classroom for the teaching of themes related to the analysis of textual structures and for identifying and selecting concepts of documental representation.

Despite the concentration of studies dedicated to the production of theoretical and methodological references to the process of document reading in subject indexing, the technique’s introspective nature favors further research into other surrounding issues, especially regarding knowledge organization bias. Due to the introspective nature of verbal protocols, their application is not limited to a specific axis, and they can contribute readily towards resolving other issues facing information professionals in multifaceted environments.

References


Kochani, A. P. (2014). *Formulação de política de indexação para coordenadorias de comunicação social em ambiente universitário: indicadores de diretrizes para análise e representação de assuntos*. Master dissertation - Universidade Federal de São Carlos, São Carlos, SP, Brazil.


Who is Tesauro? The man, words and things

Abstract

The purpose of this article is to discuss certain social and epistemic questions that seek to understand the role played by Emanuele Tesauro in knowledge organization (KO). The methodological approach taken is based on historical and epistemological reflection, aimed at understanding the elaboration and repercussions of concepts in KO. The analysis combines a diachronic historical-epistemological background with a synchronic theoretical-critical approach. It is, therefore, not a remote, historical conjecture, but a study that discusses relevant theoretical approaches to KO in historical time. Specifically, this study demarcates the relevance of Tesauro’s theory in KO. Recognizing his thinking as fundamentally a critique of language in all its distinct precepts, from logic to rhetoric, this paper indicates that Tesauro constructs a theory of discourse and a methodological process to substantiate the application of this theory, namely an index of categories and their relations. In turn, the future instrument known as the "thesaurus" responds, in part, to Tesauro’s proposition, but subverting the combinatorial range of his discursive-rhetorical influence.

1. Introduction: the shadow of a name

The purpose of this article is to discuss certain social and epistemic questions that seek to understand the role played by Emanuele Tesauro in knowledge organization (KO). The methodological approach taken is based on historical and epistemological reflection, aimed at understanding the elaboration and repercussions of concepts in KO. The analysis combines a diachronic historical-epistemological background with a synchronic theoretical-critical approach. It is, therefore, not a remote, historical conjecture, but a study that discusses relevant theoretical approaches to KO in historical time. Specifically, the study demarcates the relevance of Tesauro’s theory in KO.

It is in this context that we first pose the following research question: what “killed” Tesauro? In the 17th century, Emanuele Tesauro, author of the Aristotelian Luneta (Il Cannocchiale Aristotelico), presupposed a theory for KO in an age when critical-discourse was first emerging in a state of constant progress.

Little research in the field of information science has yet been conducted into Tesauro’s work and theory. His work has thus managed to escape the historical process anchoring the basis of KO to a whole lineage of contextual foundations. The pragmatic approach as a theoretical basis and as a theory of meaning grounded in a theory of discourse are excluded from the historical path of thought and applied practices of knowledge organization and representation.

When we enter the databases dedicated to KO, the presence of Emanuele Tesauro is practically nil. According to Saldanha and Silveira (2016), there are negligible results concerning his theories (and the apparently "curious" relation between his surname and
the revolutionary theoretical instrument that is the "thesaurus"; the instrument, not the historical person). Accordingly, there is an absence of direct and critical reflection on Emanuele Tesauro’s theory in all the bibliographic production in the context of the International Society of Knowledge Organization (ISKO).

Where, then, are there any signs within the KO field recognizing the "existence" of Tesauro? Traces of him can, in fact, be found within two main areas: firstly, within an epistemological-historical context, foreseeing certain theoretical proposals central to classification theory, but without any direct analysis of Tesauro himself or his specific approach; secondly, within an epistemological context proper, occurring at the interface between a theory of knowledge based on linguistic reflection and its unfolding into such areas as semiotics, which, in turn, are visualized at the borders between KO and linguistics (Saldanha and Silveira 2016). We can, respectively, point out the works of Almeida and Crippa (2009) and Monteiro & Giraldes (2008) as representatives of these occurrences.

However, as stated, all these are occasional and never vertical arguments, never directly dedicated to any repercussions Tesauro’s work may have had on KO. Recognizing his thinking as fundamentally a critique of language in its distinct precepts, from logic to rhetoric, we can see that the way in which Tesaurian ideas have been appropriated first overturns the notion of subject, name, and thing. Tesauro constructs a theory of discourse and a methodological process to substantiate the application of this theory, namely an index of categories and their relations. In turn, the future instrument known as the "thesaurus" responds, in part, to Tesauro’s proposition, but subverting the combinatorial range of his discursive-rhetorical influence.

2. Historical and political context: Baroque, metaphor and the 17th century

In the 17th century, an age of crises, Emanuele Tesauro formulated his theory of metaphor, the central part of his 1665 treatise on rhetoric Il cannocchiale aristotelico. It was a century characterized by religious wars, by the last great wave of the Black Plague, by the insurgency of absolutism and mercantile capitalism, and Baroque thought.

The Baroque, perceived through the prism of emerging Modernity, with its new condition of spiritual and cultural openness, contains an element that allows us to analyze a century which, at the same time, saw the birth of strong national models and international cultural thought. The element thus identified is metaphor.

Metaphor is the instrument that allows us to consider and represent the metamorphoses of the universe. It is designed to correspond to a scientific machine capable of performing a continuous transformation of culture.

It is interesting to note that the two great branches of baroque science, namely astronomy and anatomy, both share metaphorical language. Anatomy, understood as dissection and analysis, was the center of a constellation in which was found medicine,
cosmography, geography, astronomy, astrology, theology, moral philosophy, epistemology, anthropology, aesthetics, rhetoric and the fine arts.

Baroque anatomy was particularly important in the war on knowledge. Although tolerated by the Church, it was strictly controlled and limited. It was admitted only to the extent that anatomists pointed to a truth superior to scientific pursuit, recognizing an authority of spiritual power and the Truth of Christian doctrine. It spoke, therefore, of a moralized, or spiritualized anatomy.

Within the metaphorical complexity of the Baroque, there are innumerable representations of the body (or its parts), which become vehicles of moral or spiritual lessons, and anatomy becomes emblematic of the precariousness, the forfeiture, or even of the nothingness of the human condition.

Being strictly related to death, anatomy is thus the perfect emblem of our mortality (memento mori) and, at the same time, the vanity of earthly things (omnia vanitas). In all its representations, baroque anatomy becomes a spectacle, through a rich use of visual rhetoric. Such representations were to be objectively present in the construction of Tesauro’s theory of KO.

### 3. Rhetorical and “Semiotic” basis for a foundational theory of KO epistemology

The structure of Baroque thought has been met with different interpretations. One way of recognizing it lies in the notion of decay, as Carpeaux (1990) indicates. Its primary conception and application (in the field of art, mainly, but also in politics and in the theory of knowledge) indicates something strange, irregular, and out of the ordinary. At the same time, we find the notion of an accumulation of meanings and objects, of an overflowing of words and things. It is within this epistemological scenario that Emanuele Tesauro's "categorical index" and the relevance of metaphor for the classification of ideas are conceived.

According to Eco (1984), what Tesauro sought was a semantic order, based on Aristotelian categories. He focused not only on the conceptual mirror of truth behind a term, but its verisimilitude. Construction of the index involves attaching as much value to rhetoric as to logic. For Eco (1984), Tesauro’s theory constitutes one of the foundations of semiotics.

In presenting the "applied stage" of his work on organization (of knowledge, ideas and language), that is, his Categorical Index, Tesauro (1670) opens by addressing the search for a certain boundless metaphor. Here is the fundamental extremity of Tesauro’s semiotic revolution: metaphor, the aporetic limit of KO, or its complete impossibility, included as a tool and a means of considering/applying the index.

Metaphor is not, however, beyond syllogism, or beyond logic. Like Aristotle, Emanuele Tesauro (1670), integrates discursive approaches with constructs of the intellect. Thus, for him, classification includes Category (itself, in the sense of categorize, or categorie, that is, the meeting of all categories), Substance (sostanza, or
category that is above all), Quantity (quantità), Quality (qualità), Relation (relazione), Action (attione), Passion (passione), Place (sito), Time (tempo), as well as luogo (that is, the notion of space according to elements that demarcate movement) and habito (which can be interpreted as a category of contents).

Figure 1: Presentation of the Categorical Index in Emanuele Tesauro


Tesauro’s conception takes a figurative approach, including the relations between feeling, emotion and intelligence (Proctor 1973). It deals not only with the isolation and relationship between terms, but also the search to understand their effects, thus combining meanings and ideas to represent, discover and create knowledge.

After presenting the “Categorical Index”, Tesauro (1670) expands one of the central arguments of the work, the notion of sharpness or acuteness. According to Almeida & Parodi (1996), with this notion, Tesauro points to a divine property whose nature is to gather what is normally incompatible, that is, a tool uniting what is in opposition (effectively, the most “baroque” aspect of his proposition). Rhetorical figures would be the central sources for this mechanism. Tesauro, however, approaches them all through the "sharp & shrewd" figure of metaphor, which represents words themselves as concepts.

As Proctor (1973) shows, Tesauro establishes a theory of "conceit", where metaphor is central. It is a system built to classify discourse in its living, open dynamics. Therefore, its essential structure is rhetoric. Beyond theory, what we find here is a guide, a method to be applied in the development of such a system of classification by index.

4. What “killed” Tesauro? Mechanicism, Neopositivism and Political Economy

Although the name “Tesauro” resembles the Latin-derived notion (instrument) of
“thesaurus”, the historical epistemology of KO seems to conceal the theoretical and critical heritage of Emanuele Tesauro, the man.

Some potential evidence of Tesauro’s “erasure” can be historically related to social, political and epistemological issues. Various hypotheses as to why virtually all traces of him have been removed are offered here as a basis for reflection. We can thus ponder the following hypothetical-critical scenarios:

a) the "victory" of the systemic mechanicism of the Baconian method as a basis for the development of documentary languages (in view of New Organon, that is, an applied rationale based on Aristotelian logic);

b) the overriding manifestation of neopositivism in the knowledge field, fundamental both in the technical sense (the new support to mechanicism from the post-1930s electronic world) and in the theoretical sense (the support for a unity of meaning based on the syntactic-semantic dyad). The latter speaks to the "truth" behind neopositivism in terms of its autonomy. It is a way of "teaching the machine to think". We find here the essential basis of an analytical theory for the representation of concepts, that goes from Dahlberg's theory to the processes involved in conceptualizing "ontologies";

c) the political economies of science, establishing a central direction for theoretical perspectives from the most diverse fields in the 20th century, linked to industry and the market. In our context, KO is constructed as a theory applied to support developments in science and technology, responsible for allowing communication and access to hard-science products. Thus, the semantic opening offered by the theoretical-baroque position of Emanuele Tesauro is clearly a radical epistemological stance overtly at odds with the scientific-technological practices of the last century.

In summary, the historical process of the formation of ideas in KO constituted a model structured in positivist thinking and, especially, in neopositivism. This epistemological model was greatly aided by the technological and economic-political transformations which, to a large extent, removed preoccupation with the discursive (cultural) dynamics of the relations between terms, things and social groups.

5. Silent repercussions: Emanuele Tesauro in contemporary KO theory and practices

Seeking to understand Tesauro's role for KO is not a simple task, especially if we take into account the way his thought has effectively been “forgotten”. According to Saldanha & Silveira (2016), studies on Tesauro’s method in KO are rare and superficial.

In critical-historical works, such as Dahlberg (1993) and Hjorland (2008), Tesauro's point of view is not considered. It is not even mentioned in recent reviews, such as Clarke (2017).

Historically, the main references for classification studies within this context have touched on the questions contained in Tesauro’s theory. An example of this are the categorization, factioning, and indexing processes popularized at the heart of KO by
such names as Ranganathan (1967), Vickery (1980), Dahlberg (1978) and Lancaster (1993), as well as members of the Classification Research Group (CRG).

When we analyze the influence of Aristotle on the 20th century authors of KO, in the notion of concept and the facet analysis method, for example, we can find clear traces of Tesauro. We can see that his notion of the "categorical index" is present nowadays. The main evidence, however, comes not only from Aristotelian thinking. It is also possible to perceive a specific way of analyzing language in the ordering of words and things.

In other words, even in the neopositivist tradition, the influence of Aristotle can also be seen in Emanuele Tesauro. On the other hand, when we understand the urgent culturalist approaches focusing on the contextual languages of each community (that is, contemporary approaches that focus on the social and cultural issues of knowledge representation), we sense the force of a theory of metaphor for KO. In this way, Tesauro's point of view has now been objectively reinstated. His metaphor theory places the dynamics of language at its heart. Therefore, the social and cultural transformations manifested in language can be identified, discovered and created from their “categorical index”.

6 Final remarks: who is Tesauro?

The term "thesaurus", with reference to Emanuel Tesauro’s name, can be interpreted in different ways. In epistemological terms, we can call the "thesaurus" a "discourse theory for knowledge organization". In a methodological and artefactual approach, "thesaurus" means "discursive index". The two ways of delineating the meanings of "thesaurus" based on the historical personage of Tesauro are already of sufficient complexity as to allow for a broad analysis in the field.

In the first case, we are arguing that Tesauro presupposes an epistemological-historical premise rooted in the philosophy of language; a radical context, capable of perceiving both the logical and rhetorical dynamics of knowledge representation. The notion does not, therefore, exclude even part of the neopositivist advances made after the 1930s, but it demonstrates their limitations for solving problems in KO.

In the second case, we are highlighting the fact that the "thesaurus" instrument (the same one currently being developed for direct and local solutions, such as languages for a given library or documentation center, or for complex and intersubjective solutions, such as a thesaurus of education, or a thesaurus of Afro-descendant cultures) is already present in the theoretical and applied argumentation of Emanuele Tesauro's "categorical index", and, therefore, it can also be seen as being nominally linked to Tesauro, the man.

Tesauro’s theory, under the influence of Aristotle, seeks to investigate the meaning dilemmas at the lexical frontiers between knowledge domains. We can thus conclude that Tesauro is one of the most important founders of KO theories.
Acknowledgment

This research was developed under the financial support of the National Council for Scientific and Technological Development (CNPq), Brazil, and the Carlos Chagas Foundation for Research Support of Rio de Janeiro State (FAPERJ), Rio de Janeiro, Brazil.

References


Wikipedia categories in research: towards a qualitative review of uses and applications

Abstract
This paper reviews the study and use of the Wikipedia category system in research, adopting a systematic literature review approach, while conducting a qualitative review of the content from a set of selected papers. Several types of research are identified, from intrinsic studies of this category system to analysis of its use as a tool for analyzing documentary corpora other than Wikipedia. We conclude that the Wikipedia system of categories offers a valid classification scheme for the different approaches taken to study knowledge organization in multiple contexts.

1. Introduction
Wikipedia is the primary resource of encyclopedic information available on a global scale. With more than 32 million articles, it is consulted by millions of users. It can be considered as a knowledge base, structured and labeled according to basic instructions and parameters. Looking at developments in structuring data according to semantic criteria, then the most significant existing knowledge base is the Wikidata project (Vrandecic 2013). By analyzing the structure of articles in Wikipedia, as well as the tools for organizing and exploring the encyclopedia, we can determine that among its key elements, inherent to its content, are categories.

The set of categories, which comprises a vast collection of constantly evolving terms, is used by editors to frame the contents within a structure of knowledge organization (Capocci, Rao & Caldarelli, 2008). Categories were introduced in Wikipedia in 2003 and pages of categories and subcategories in 2004. Both elements are collaboratively defined, maintained and updated by the community (Thornton and MacDonald 2012). It is a system that combines a hierarchical organization with relations between different categories giving rise to the creation of polyhierarchies and associations. Effectively, the category system is a system of the body of knowledge, and it can group items into sets and conceptual or topical subsets.

The scientific study of Wikipedia’s dynamics has resulted in the publication of papers on collaborative editing processes, behavioral patterns of user communities, vandalism, etc. (Mesgari et al. 2015). The undertaking of research into the universe of Wikipedia has enabled the development and availability of several systematic literature reviews, catering for different approaches (Tramullas 2015). Research about Wikipedia shows different orientations, objectives, methods, and results. Nonetheless, it is also possible to identify in the bibliography a set of studies, which are using the textual corpus of Wikipedia. Some of them use categories as a parameter for different analysis
related to organization and information retrieval (IR), classification of documents, semantics and ontologies, or social labeling.

In line with all these aspects, the main objective of this paper is to identify how researchers are using and applying the Wikipedia category system and how this purpose is being reflected in the evolution of the corpus of scientific literature available. Secondly, it aims to review how a knowledge organization system, developed collaboratively, is being used as a research resource in different approaches to the treatment and organization of information, and what authenticates its validity as a classification system. Finally, it should be noted that this work does not intend to assess the structure, evolution or the quality of the Wikipedia category system.

2. Methodology

The methodology used to carry out this study was a systematic bibliographic review. In this case, we opted for a qualitative study, selecting specific papers to review, rather than a quantitative descriptive research or bibliometric approaches. The compilation of bibliographic data was carried out through reference queries into the topic under investigation conducted in Web of Science and Scopus. The working method adopted was proposed by Okoli and Schabram (2009) for the study of research on Wikipedia and has already been previously applied by the authors (Tramullas, Garrido-Picazo and Sánchez-Casabón 2016). During the first phase, the sources and the consultation to perform were selected. Queries on Scopus and Web of Science were held between November 2017 and January 2018, using the query “Wikipedia and categories,” the search being limited to papers published between 2002 and 2017. 666 results were obtained in Scopus and 311 in Web of Science. In both cases, the first results identified were in 2006. Due to the objectives and limits of this work, neither the ACM nor the IEEE digital libraries were consulted for repeated content. Google Scholar was likewise not used, due to the impossibility of limiting searches to the specific positions of documents or their bibliographic references.

In a second phase, once we had obtained the raw data from the references, we processed the datasets. First, both datasets were merged, to continue with the identification and elimination of duplicate documents. This task reduced the number of papers to 680. Subsequently, a qualitative selection of the works was carried out, considering the thematic content identification of the different studies, by reviewing the titles, abstracts, and author keywords assigned to the same. The selection criteria established was the use or study of categories as an essential element of the work reviewed in each case. Each paper was considered by three authors independently. To avoid discrepancy, each article was examined to determine its inclusion or rejection by a majority. This type of selection does not allow for an assessment of quality, although this criterion was not taken into account since this study neither aims to elaborate a selected bibliography nor to evaluate the quality of the contributions or their impact.
The filtering carried out made it possible to eliminate those articles or communications from the set whose use or reference to Wikipedia categories was not directly related to the objectives of this study. Finally, the number of selected works amounted to 546. Studies and selected papers were reviewed to identify the use made in them of the Wikipedia category system. We have defined the application context, the method used and the results obtained, to delineate the studies which have used Wikipedia categories as an integral part of their research. Finally, we proceeded describe the results and the develop a synthesis as well as define proposals. All bibliographic references collected and reviewed in the study have been published in open access through specific groups in Mendeley and Zotero¹.

3. Results

The results obtained from the qualitative review carried out demonstrate the variety of ways in which researchers have approached, used, and applied the Wikipedia category system. This wealth of material, in turn, implies a limit for the qualitative analysis proposed: the combination of techniques, approaches, and applications existing in research work makes it impossible to establish precise divisions among types of proposals. While we may find studies ascribed to a subject, for example, generation of ontologies, there are other works that also generate ontologies but in combination with other information retrieval (IR) techniques. Such research can take place in a generic context or be confined to specific domains. Accordingly, any qualitative treatment such as that addressed in this study must be limited to delineating identified topics and research areas. To investigate the relationships between issues and possible divisions or types included, it would be necessary to apply information processing techniques based on numerical analysis of the bibliographic corpus, as recently developed by Smiraglia and Cai (2017).

In response to the limitation indicated above, the qualitative review enables an initial division to be made. Firstly, studies that analyzed the category system itself within the context of Wikipedia (covering aspects such as organization and lifecycle of content, structure, user community perspective, or the evolution and improvement of the category system). Secondly, those papers that use Wikipedia categories in the context of studies on different aspects of information processing, usually using documentary corpora outside Wikipedia (collections of documents of different types and thematic, web pages, messages on social networks, etc.). About 85-90% of the studies reviewed belong to this second group. It is necessary to emphasize the presence of studies that use corpus created ex-profeso, that is, which have been generated or extracted from Wikipedia itself. All the studies reviewed could be framed in any of the four research

categories on Wikipedia listed by Nielsen (2017), although this researcher places the subject treated in this proposal within the generic category of "Wikipedia uses".

Within this large group of works, several large subdivisions can be made, although as previously indicated some studies could be attributed to two or more of them. Proposed groups correspond to:

1. Information retrieval: those proposals which use categories in different processes and techniques of IR, both as regards the formulation of search expressions, their refinement, and improvement, or to the filtering of results. IR should be highlighted both for its use in performance evaluation as well as recommendation processes.

2. Entity processing: for particular interest in quantity, an emerging area of analysis that seeks to identify (named) entities in documents. Such studies have been extensively applied since they serve to identify semantic relationships among terms, solve disambiguation, or integrate classifications and taxonomies. Entity processing may be applied to research on natural language processing and even the development of dictionaries in different languages or multilingual dictionaries.

3. Indexing and classification of document corpus: categories or specific subsets are tools to proceed with the indexing and classification of document sets, usually within particular contexts or domains. A subset of these works is made up of those which use categories to label textual documents, within the framework of automatic indexing processes. Another subset to highlight is that formed by studies that deal with the automatic labeling of video or photography. This may include studies that produce specialized corpus automatically, using Wikipedia categories in turn.

4. Creating and using taxonomies: it is one of the most classic uses. Wikipedia categories are extracted from their context to form new schemes, applied in specific domains. They can be completed with the combined use of other taxonomies. In relation to this approach, we have identified some papers that propose creating classifications of classical structure, as hierarchical classifications or thesauri. On numerous occasions, taxonomies created are integrated into processes corpus indexing and rating and collected in the previous point.

5. Creating and using ontologies: the second of the traditional uses. About 15% of the works reviewed deal with the creation and use of taxonomies and ontologies from the Wikipedia category system. As taxonomies indicated in point 4, they are used in document classification processes, but also in ontological engineering and development of semantic relationships between entities.

6. Semantic treatment: this group included different approaches which characterize the principles and techniques of the semantic web, methods such as the creation
of category graphs, trees and schemata, extraction of triplets, identification of meaningful relationships among terms and their semantic use, etc. Ontologies, although an integral part of the semantic web, have been included in a separate group because of their importance.

7. Other uses: within this group can be included very specific applications or those which are numerically underrepresented in the studies reviewed. Examples of these may include user profiling, collaborative editing patterns, or identification of events and people.

In order to identify more precisely the topics that have been the subject of research, we have developed a list of significant terms and expressions, extracted from the titles and abstracts reviewed (table 1).

Table 1: Significant expressions identified

<table>
<thead>
<tr>
<th>Information Retrieval</th>
<th>Entity Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Automatic question generation</td>
<td>• Query expansion for entity ranking</td>
</tr>
<tr>
<td>• Automatic subject induction</td>
<td>• Query phrase expansion</td>
</tr>
<tr>
<td>• Discovering answers</td>
<td>• Semantic question answering</td>
</tr>
<tr>
<td>• Entity retrieval</td>
<td>• Supervised question classification</td>
</tr>
<tr>
<td>• Exploratory search</td>
<td>• Tagging</td>
</tr>
<tr>
<td>• Organizing search results</td>
<td>• Wikipedia categories for ad hoc search</td>
</tr>
<tr>
<td>• Query classification</td>
<td></td>
</tr>
</tbody>
</table>

|                                                                                      |                                                                                   |
| Indexing and classification of documentary corpus                                    |                                                                                   |
| • Automatic document classification                                                  | • Matching named entities                                                         |
| • Automatic document tagging                                                         | • Named entity extraction                                                         |
| • Categories for document labelling                                                  | • Named entity linking                                                            |
| • Conceptual indexing                                                                | • Query expansion for entity ranking                                              |
| • Corpus building in machine learning                                              | • Semantic tags                                                                    |
| • Document clustering                                                                | • Wikipedia entity expansion                                                      |
| • Document context similarity                                                        | • Word sense disambiguation                                                       |

|                                                                                      |                                                                                   |
| Creating and using taxonomies                                                       |                                                                                   |
| • Analysis of cluster structure                                                     | • Taxonomy-based information content                                              |
| • Automatic taxonomy extraction                                                     | • Twixonomy                                                                       |
| • Comparing taxonomies                                                               | • Web taxonomies                                                                   |
| • Derivation of “is a” taxonomy                                                     | • Wikipedia category graph                                                        |
| • Method for refining a taxonomy                                                    |                                                                                   |
| • Taxonomy and clustering                                                            |                                                                                   |
While the variety of terms used confirms and reinforces the results indicated in the qualitative review, it does, however, highlight an underlying problem with systematic reviews, namely the disparity of author opinions in the drafting of titles, abstracts and selecting keywords. Even in some cases, we detect the use of synonyms of terms or expressions that are not. Regarding the classification and identification of the content of the works with greater precision, the results show that researchers resort to mixed approaches and combine methods and techniques, hindering a traditional plan, and it requires methods of automatic processing of information for better results.

4. Conclusions

It must be admitted that Wikipedia is having a significant influence on the way that users approach the resolution of their information problems, and a significant positive impact (Fallis, 2008). Scientific research is not oblivious to the importance of this information resource (Tomaszewski and MacDonald 2016). It is not unusual to research into the organization and retrieval of information and find in Wikipedia a high-value

<table>
<thead>
<tr>
<th>Creating and using ontologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automated construction of domain ontology taxonomies</td>
</tr>
<tr>
<td>Automatic ontology generation</td>
</tr>
<tr>
<td>CyC ontology</td>
</tr>
<tr>
<td>Deriving domain taxonomies</td>
</tr>
<tr>
<td>Domain ontological structure</td>
</tr>
<tr>
<td>Evolution of ontologies</td>
</tr>
<tr>
<td>Extracting ontologies</td>
</tr>
<tr>
<td>Mining concepts</td>
</tr>
<tr>
<td>Ontological models</td>
</tr>
<tr>
<td>Ontology density</td>
</tr>
<tr>
<td>Ontology evaluation</td>
</tr>
<tr>
<td>Rich ontology extraction</td>
</tr>
<tr>
<td>Semi-automatic ontology creation</td>
</tr>
<tr>
<td>Topic ontology</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Semantic treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category annotation recommendation</td>
</tr>
<tr>
<td>Category graph</td>
</tr>
<tr>
<td>Concept graph</td>
</tr>
<tr>
<td>Disambiguation of keyword search results on highly heterogeneous structured data</td>
</tr>
<tr>
<td>Domain semantic networks</td>
</tr>
<tr>
<td>Entity linking</td>
</tr>
<tr>
<td>Knowledge trees</td>
</tr>
<tr>
<td>Ontology density</td>
</tr>
<tr>
<td>Ranking related entities</td>
</tr>
<tr>
<td>Semantic knowledge base</td>
</tr>
<tr>
<td>Semantic knowledge extraction</td>
</tr>
<tr>
<td>Semantic relatedness</td>
</tr>
<tr>
<td>Semantic relationships extraction</td>
</tr>
<tr>
<td>Semantic resource extraction</td>
</tr>
<tr>
<td>Semantic tagging</td>
</tr>
<tr>
<td>Semantically related category hierarchies</td>
</tr>
<tr>
<td>Triples extraction</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Other uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic blog classification</td>
</tr>
<tr>
<td>Automatic mapping of Wikipedia categories</td>
</tr>
<tr>
<td>Automatic thesaurus generation</td>
</tr>
<tr>
<td>Concept hierarchies</td>
</tr>
<tr>
<td>DBpedia</td>
</tr>
<tr>
<td>Dynamic facet hierarchy construction</td>
</tr>
<tr>
<td>Generation of dictionaries</td>
</tr>
<tr>
<td>Multilingual domain specific resources</td>
</tr>
<tr>
<td>Semantic interest profiles</td>
</tr>
<tr>
<td>Semantic knowledge base</td>
</tr>
<tr>
<td>Semantic recommender</td>
</tr>
<tr>
<td>Semantic tags</td>
</tr>
<tr>
<td>Terminology</td>
</tr>
<tr>
<td>Wikipedia categories clustering</td>
</tr>
<tr>
<td>WordNet</td>
</tr>
<tr>
<td>YAGO-NAGA</td>
</tr>
</tbody>
</table>
The first conclusion to draw from this study is that Wikipedia is a hot topic for different research fields, both in its internal aspects and external use of its data in other areas and research approaches. Secondly, the Wikipedia category system effectively constitutes the socio-temporal evolution of knowledge organization processes in collaborative environments. Thirdly, it is necessary to emphasize its use as a tool of support and validation in different types of approaches to the study and analysis of documentary corpora.

When dealing with Wikipedia articles and its constantly evolving documentary corpus category system, we note that results obtained in different studies may vary, in the medium or long term, according to the development of external and internal factors to the encyclopedia itself. For example, as a reference we took social labeling systems, an area of great interest during the last decade which have been gradually disappearing from research on knowledge organization.

Finally, we can clearly emphasize the potential that the Wikipedia category system offers, as it is a universal classification scheme developed collaboratively, in contrast to functional classification schemes developed in closed contexts. It provides a broad field both for schema validation and for creating new schema from a perspective that allows a combination of both approaches to knowledge organization. Several recent papers address this issue (Salah, Gao, Suchecki y Scharnhorst, 2012; Kiyota et al. 2009), as well as the advantages and advances of the Wikipedia system versus traditional classifications (Jimenez-Pelayo 2009).

References


Interoperability towards Information Access
A Comparative analysis and evaluation of bibliographic ontologies

Abstract
This paper evaluates two bibliographic ontologies developed by computer scientists alone, without the assistance of knowledge organization (KO) scholars: the Bibliographic Ontology (Bibo) and FaBiO (FRBR-aligned Bibliographic Ontology). The aim of the study is to analyze their classes and properties, evaluate the use of categories in Bibo, verify the alignment with FRBR in FaBiO, and check the consistency of the definitions of classes with categories belonging to the cataloging field. Finally, the study evaluates the properties presented in Bibo and FaBiO, checking them against the properties in Bibframe 2.0. The paper presents tables of comparison between FaBiO classes and FRBR categories, also considering the IFLA-LRM high-level conceptual model. Moreover, it presents a mapping between FaBiO and Bibo properties and Bibframe 2.0 properties, highlighting the fact that the former have fewer properties than the latter.

Introduction
Ontologies play an essential role in building linked open data (LOD) to enhance the Semantic Web. They offer a tool to express semantically qualified relationships in the realization of RDF triples and the possibility of linking data from different datasets.

As Tim Berners-Lee (2006) has asserted, at the highest level the process of LOD creation allows connection of data within the Web to enrich information by interlinking structured data from different sources. Ontologies are used to define suitable and qualified relationships in the process of realizing RDF triples, and, as a result, to improve semantic interoperability. Among the ontologies used in the process of interlinking, FOAF\(^1\) enables definition of biographic profiles and relations among persons and groups, and the Organization ontology\(^2\) allows expression of organizational structures, including governmental institutions. It would be a good strategy to use vocabularies and ontologies well known by the Web community, rather than to develop new ontologies.

Bibliographic ontologies enable description of entities that belong to the bibliographic set, such as textual publications (e.g., articles, books, and series) and web pages, datasets, films etc., and define the relationships between these bibliographic entities. Apart from autographed manuscripts and single-edition works, in most cases bibliographic entities in library collections present a complex set of relationships. For instance, “derivative relationships” concern different editions of the same work, translations in different languages, or works derived from a pre-existing work, and “sequential relationships” include sequels of a monograph (Tillet 1989; Green 2001).

\(^1\) http://www.foaf-project.org/
\(^2\) http://www.w3.org/TR/vocab-org/W3C Recommendation 16 January 2014.
A work may be a part of, the logical continuation of, or the transformation of another work (IFLA 2017, 69-78).

Bibliographic ontologies should define specific relationships, such as authorship, editorship, and aboutness among the entities, as well as the relationships that connect works and their abridgments, adaptations, and translations. In addition, bibliographic ontologies can underline relationships between a serial and the transformations it may have had over time, such as supplements or successors.

Bibliographic data models such as Bibframe⁢³ FRBRoo⁣⁴ and Dublin Core Metadata Element Set⁵ are well known and well founded on bibliographic expertise. Additionally, there are other bibliographic ontologies available on the Web, made by computer scientists alone or scholars from outside the knowledge organization (KO) field. These require careful consideration from scholars dealing with bibliographic entities and relationships. Among the challenges and opportunities for KO in the digital age is control over the increase of bibliographic ontologies. Scholars in KO should take interest in bibliographic ontologies, monitor and scientifically supervise the development of new ontologies, considering the internationally accepted conceptual models developed in the field of KO.

Objectives

This paper seeks to analyze and evaluate two bibliographic ontologies developed without the contribution of KO scholars. In particular, the paper analyses Bibo (The Bibliographic Ontology)⁶ and FaBiO (FRBR-aligned Bibliographic Ontology)⁷. The aim of the study is to analyze the classes and properties defined in these two ontologies, evaluate the use of categories in Bibo, and verify the alignment with FRBR in FaBiO. Finally, it checks the consistency of definitions of classes with categories belonging to the cataloging field. Moreover, the study endeavors to evaluate the properties provided by Bibo and FaBiO, checking them against the properties offered by Bibframe 2.0.

Bibo and FaBiO

Bibo was developed in 2009 by Bruce D’Arcus and Frédérick Giasson, and it is the first OWL ontology written in RDF that provides main concepts and properties for describing bibliographic entities and citations. The BNB Linked Data Platform, which provides access to the British National Bibliography published as linked open data, has

---

³ https://www.loc.gov/bibframe/.
⁵ http://dublincore.org/documents/dces/.
used Bibo properties since 2011, along with the Dublin Core Metadata Element Set. The Linked Data Service of the Deutsche Nationalbibliothek has also used Bibo and Dublin Core terms since 2010.

The developers of FaBiO, Silvio Peroni and David Shotton, assert that it is an ontology based on FRBR (IFLA 1998, 2009) entities and that it was created for describing entities from different and correlated points of view on the Semantic Web. The FaBiO Work class is restricted to published or printable entities: textual publications, such as articles, books, series, and journals, etc. The entities also include web pages, datasets, computer algorithms, catalogs, etc. FaBiO is part of SPAR (Semantic Publishing and Referencing Ontologies)\(^\text{10}\) a set of complementary and orthogonal ontologies developed in OWL 2 DL\(^\text{11}\) by Bologna University and Oxford University (2009). SPAR participates in the Semantic Publishing project, which deals with the use of Semantic Web technologies with the aim of semantically linking scientific literature to facilitate its discovery. SPAR ontologies “describe the different aspects of the scholarly publishing domain”\(^\text{12}\).

At present, FaBio includes seven super-classes, one equivalent class, and 237 subclasses (for a total of 245 Classes), 28 object properties\(^\text{13}\), 65 datatype properties\(^\text{14}\) and 15 named individuals. Bibo, in comparison, includes five principal classes and 34 subclasses, 32 object properties and 20 sub-properties, 20 datatype properties (10 of which are OWL equivalent properties) and 26 sub-properties.

**Methods**

Recently, Terhi Nurmikko-Fuller et al. (2016) analyzed five bibliographic models, Bibframe, Mods, Mads, Schema.org e FRBRoo, with the aim of finding the differences and similarities in these models. The mapping they provided highlights five types of alignment for the use of classes and properties.

In the present study, a comparative methodology has instead been proposed with the

---

\(^{8}\) http://www.bl.uk/bibliographic/pdfs/publishing_bnb_as_lod.pdf.

\(^{9}\) http://www.dnb.de/EN/Service/DigitaleDienste/LinkedData/linkeddatalod_node.html.

\(^{10}\) http://purl.org/spar.


\(^{12}\) The SPAR ontologies are twelve in all: FRBR-aligned Bibliographic Ontology (FaBiO); Citation Typing Ontology (CiTO); Bibliographic Reference Ontology (BiRO); Citation Counting and Context Characterisation Ontology (C4O); Document Components Ontology (DoCO); Publishing Status Ontology (PSO); Publishing Roles Ontology (PRO); Publishing Workflow Ontology (PWO); Scholarly Contributions and Roles Ontology (SCoRO); DataCite Ontology (DataCite); Bibliometric Data Ontology (BiDO);Five Stars of Online Research Articles Ontology (Five).

\(^{13}\) They are organized into the following groups: Top Object Property, Has embodiment, Has exemplar, Has subject term, Related endeavor, Has part, Has realization, Is embodiment of, Is exemplar of, Is part of, Is realization of, Is representation of, Is scheme of.

\(^{14}\) They are organized into the following super-properties: Top data properties, Has title, Has identifier, Has date.
aim of making an analysis and a reliable assessment of the classes and properties of Bibo and FaBiO. The definitions of the classes in Bibo and FaBiO are compared with the bibliographic categories used in the cataloging field that provides the framework within which the entities should be evaluated. In particular, the FaBiO classes are related to the FRBR model and to the recently published IFLA – LRM (Library Reference Model), which is the synthesis of the FR family of conceptual models previously and separately published, such as the Functional Requirements for Authority Data (FRAD) and the Functional Requirements for Subject Authority Data (FRSAD). LRM provides principles for bibliographic information and can be used in LOD building for improving the use of bibliographical data in the LOD environment.

Moreover, an examination of the properties exposed in FaBiO and Bibo has been carried out comparing them with the vocabulary of Bibframe 2.0\textsuperscript{15} provided by The Library of Congress. In particular, the Cataloging Resource Relationships (general, specific and detailed),\textsuperscript{16} mainly sub-properties of the property relatedTo, are taken into consideration.

### Analysis and discussion

Bibo presents a variety of entities related to the bibliographic world, organized into five principal classes: Agent, Collection, Document, Document status, Event, and 34 subclasses. Bibo asserts that the classes Agent and Document (and the subclass Image) are equivalent to FOAF corresponding classes. The classes Document and Collection accommodate the majority of the bibliographic sub-entities. Among the subclasses of Document, we find Article, Book, Image, Legal Document, Manuscript, Report, Web page, etc. Among the subclasses of Collection are Periodical, Series, and Web site.

FaBiO has been developed for describing and publishing any bibliographic entity on the Semantic Web, prioritizing textual entities, such as books, articles, and journals. It provides classes relevant for academia.

The FaBiO classes have been structured using the FRBR categories: Work, Expression, Manifestation, and Item. The objective is to describe a document considering its different Expressions. For instance, an academic paper could firstly be published as a journal article, later as a paper in conference proceedings, or as a book chapter (Peroni and Shotton 2012, 36). The developers intentionally aligned FaBiO with FRBR categories to permit greater expressivity (Shotton 2011). In fact, FaBiO enables the definition of the various formats of an academic article. FaBiO allows creation of machine-readable RDF metadata and easily imports terms from DC, FRBR,

\textsuperscript{15} https://www.loc.gov/bibframe/.
\textsuperscript{16} http://id.loc.gov/ontologies/bibframe-category.html.
FRBR Core\textsuperscript{17}, SKOS\textsuperscript{18} and PRISM.\textsuperscript{19} Entities such as “contributor” and “creator” in FaBiO are considered as annotation properties, which are mainly imported from DC or RDFS.

In the present survey, I have been focusing on the evaluation of the use of FRBR categories in FaBiO to distinguish different aspects of the publication process. The FaBiO classes have been compared with the FRBR categories (Work, Expression, Manifestation, and Item) also considering the IFLA-LRM high-level conceptual model.

As a result, the mapping between relevant classes in FaBiO on one hand, and the categories and the definitions provided by FRBR and IFLA-LRM on the other, emphasizes that alignment is restricted to a simplistic level and is not always correct.

Table 1 shows a sample of categories used in FaBiO (column 1), definitions in FRBR (column 2), and categories used in IFLA-LRM (column 3).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Abstract</strong></td>
<td>Expression</td>
<td>“Abstracts, digests and summaries [...] represent new works” (p. 18)</td>
<td>Work (p. 21)</td>
</tr>
<tr>
<td><strong>Article</strong></td>
<td>Expression</td>
<td>Component/part of a W. (p. 67-68)</td>
<td>Work (p. 39)</td>
</tr>
<tr>
<td><strong>Book</strong></td>
<td>Expression</td>
<td>Format of Manifestation (p. 21)</td>
<td>-</td>
</tr>
<tr>
<td><strong>Critical edition</strong></td>
<td>Work</td>
<td>-</td>
<td>Expression (with critical notes) (p.23)</td>
</tr>
<tr>
<td><strong>Film</strong></td>
<td>Movie (Expres.)</td>
<td>-</td>
<td>Work (p. 21)</td>
</tr>
<tr>
<td><strong>Index</strong></td>
<td>Expression</td>
<td>Referential Work (p. 63-64)</td>
<td>-</td>
</tr>
<tr>
<td><strong>Supplement</strong></td>
<td>Expression</td>
<td>Referential Work/Express.; Autonom. W/E (p. 63-73)</td>
<td>Work/Expression (p. 70)</td>
</tr>
</tbody>
</table>

It is important to notice that in FaBiO the distinction between Work and Expression is not clear enough. FRBR and the recent IFLA-LRM have defined levels of bibliographic description and have established a strict connection between Work, Expression, Manifestation and Item: “The Work consists of the intellectual or artistic creation that lies behind all the various Expressions of the Work” (IFLA-LRM 20). A Manifestation is a Manifestation of an Expression of a Work. FaBiO, instead, uses the FRBR categories with the aim of defining the different types of bibliographical entities or objects that belong to the publishing world. The relationships between Manifestations, Expressions and Work are lost. In the hierarchical organization of

\textsuperscript{17} http://vocab.org/frbr/core.
\textsuperscript{18} http://www.w3.org/2004/02/skos/.
\textsuperscript{19} http://www.prismstandard.org/specifications/.
classes and subclasses, among the subclasses of Work we find not only artistic work, critical edition, essay, image, reference work, report, research paper, review, and vocabulary; but also dataset, metadata, and grant application. Among the subclasses of Expression are abstract, article, book, brief report, chapter, comment, conference poster, index, letter, manuscript, metadata document, movie, periodical issue, proceedings paper, report document, repository, and supplement, as well as web content, computer program, database, and e-mail.

It is not clear what the rationale is for classifying an entity as Work – for example, essay, report, research paper, review; or as Expression – abstract, article, manuscript, brief report, chapter, or proceedings paper. In many cases, the entities classified as Expression should have been classified as Work. In my opinion, the developers of FaBiO confuse the FRBR bibliographic categories (Work and Expression) with types of publications, such as books, articles, or issues. It is also to be noted that alignment with FRBR categories in the hierarchical organization of FaBiO subclasses is much more complex. For instance, Review is classified as Work; Book Review is a subclass of Review (Work); Review Article is a subclass of Article (Expression).

The properties provided by Bibo and FaBiO are now analyzed and compared with Bibframe 2.0 properties. Bibframe 2.0 presents “three core levels of abstraction: Work, Instance, and Item” and “additional key concepts that have relationships to the core Classes”20, such as Agents, Subjects, and Events (associated with Works or Instances). The Bibframe vocabulary21 offers 75 classes and 112 subclasses (plus two FOAF classes), 194 properties, out of which 131 object properties (and sub-prop.), and 63 datatype properties (and sub-prop.). FaBiO presents a small number of object properties (28); Bibo, on the other hand, presents a greater number of these (52), versus the 194 properties in total provided by Bibframe 2.0.

Table 2 offers a mapping between all the object properties presented in FaBiO, part of the object properties offered in Bibo, and the properties of Bibframe 2.0 pertaining to the bibliographic field. Table 2 presents only 18 out of the 52 Bibo object properties, as the majority of these are generic or not strictly speaking devoted to bibliographic relations (event, agent of event, place of event, etc.). FaBiO offers mainly object properties that follow FRBR alignment and correspond only marginally with Bibframe properties, due to the reduction in core classes, as Bibframe provides only Work, Instance, and Item. Besides, FaBiO adds three properties, and their inverse properties: hasManifestation, hasPortrayal, hasRepresentation.

Moreover, the mapping offered in Table 2 highlights the fact that FaBiO and Bibo lack properties concerning derivative, merging, and absorbed resources, which are provided instead by Bibframe. Bibframe offers sub-properties of the property

---

20 https://www.loc.gov/bibframe/docs/bibframe2-model.html.
relatedTo, such as accompaniedBy, to define a supplement or index added to a resource, and derivativeOf, to express translations and different editions of a work. Additionally, it provides the property precededBy to describe the replacement of a resource with another, the merging of two or more resources to form a new resource, the continuation of a resource under a new title, and the incorporation of a resource into another, and succeededBy to define a resource that supersedes another and the division of a resource into two different resources. These properties are highly relevant to connect entities that belong to the bibliographic field; nevertheless, they are not present in Bibo and FaBiO.

Table 2: Mapping between object properties in FaBiO, Bibo, and Bibframe 2.0.

<table>
<thead>
<tr>
<th>1. FaBiO object properties</th>
<th>2. Bibo object properties</th>
<th>3. Bibframe 2.0 object properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relation (reviewOf)</td>
<td>review</td>
<td></td>
</tr>
<tr>
<td>cites/citedBy</td>
<td>references</td>
<td></td>
</tr>
<tr>
<td>isReferencedBy</td>
<td>referencedBy</td>
<td></td>
</tr>
<tr>
<td>distributor</td>
<td>acquisitionSource</td>
<td></td>
</tr>
<tr>
<td>owner</td>
<td></td>
<td></td>
</tr>
<tr>
<td>reproducedIn</td>
<td>reproductionOf</td>
<td></td>
</tr>
<tr>
<td>has creator, has responsible contributor (author, edit.) agent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>has discipline</td>
<td></td>
<td></td>
</tr>
<tr>
<td>has embodiment /is embod. of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>has exemplar / is exemplar of</td>
<td>hasItem / ItemOf</td>
<td></td>
</tr>
<tr>
<td>has format</td>
<td>format</td>
<td></td>
</tr>
<tr>
<td>has language</td>
<td>language</td>
<td></td>
</tr>
<tr>
<td>has license</td>
<td></td>
<td></td>
</tr>
<tr>
<td>has manifestation/ is manif. of</td>
<td></td>
<td>hasInstance / instanceOf</td>
</tr>
<tr>
<td>has part</td>
<td>hasPart</td>
<td>hasPart (seriesOf, subseriesOf) / PartOf</td>
</tr>
<tr>
<td>has portrayal/ is portrayal of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>has publisher</td>
<td>publisher (issuer)</td>
<td></td>
</tr>
<tr>
<td>has realization/ is realiz. of</td>
<td>isVersionOf (transl.Of) instanceOf / hasInstance</td>
<td></td>
</tr>
<tr>
<td>has representation/ is repr. of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>has rights</td>
<td>rights</td>
<td></td>
</tr>
<tr>
<td>has subject term</td>
<td>subject</td>
<td>subject</td>
</tr>
<tr>
<td>is discipline of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>is in scheme</td>
<td></td>
<td></td>
</tr>
<tr>
<td>is part of</td>
<td>isPartOf (reproducedIn)</td>
<td>part of (hasSeries, hasSubseries)</td>
</tr>
<tr>
<td>is scheme of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>is stored on</td>
<td></td>
<td></td>
</tr>
<tr>
<td>stores</td>
<td></td>
<td></td>
</tr>
<tr>
<td>accompanies (issuedWith, supplementTo)</td>
<td>accompaniedBy (supplement, index)</td>
<td>derivativeOf (translationOf, OriginalVersion)/hasDerivative</td>
</tr>
</tbody>
</table>
It is also important to bear in mind that, even though Bibframe 2.0 gives special attention to derivative and added and merged resources, it should also have paid full attention to the properties that belong to continuing resources, such as serials. The formal ontology PRESSoo (2016), a recently developed extension of FRBRoo\textsuperscript{22}, addresses the problems of absorption, continuation, replacement, separation, and merging, also provided by Bibframe 2.0. Moreover, PRESSoo adds, among others, the cases of temporary replacement of a serial with another serial, the reprint of a dead serial such as a monograph, the enhancing of a series by a monograph, the launch and end of a periodical, issuing rules (for instance, regularity, frequency, etc.), and the partial continuation of a serial.

In addition, it is worth mentioning that in Bibframe 2.0 the property DerivativeOf/hasDerivative should have been better defined. This property should include, for instance, the following cases: a Work that derives from another Work; a Work that is a continuation of another Work, abridgements and adaptations (see FRBR), and an Expression that derives from another Expression of a Work.

**Conclusions**

This survey highlights the problems that arise in developing a formal ontology in the bibliographic field, available for the Web community, and in modelling classes, subclasses and properties. The study emphasizes that the two bibliographic ontologies analyzed, FaBiO and Bibo, are inadequate from many points of view. Even though the developers of Bibo curated, in particular, the soundness of the definitions of the classes that belong to the bibliographic field, this ontology is poor with respect to the properties required in a bibliographic context. The FaBiO ontology provides a simplistic alignment to FRBR categories, not grounded in sound cataloging principles, and thus inadequate to reinforce the ontology. In addition, both ontologies show insufficient attention to bibliographic relationships concerning derivative or merged resources.

The evaluation of these two bibliographic ontologies highlights the fact that to realize the process of interlinking different datasets, it is necessary to make reference

to bibliographic data models provided with the collaboration of researchers in the cataloging field. It stresses the need for assistance from KO scholars in the modelling of bibliographic ontologies and control over the growth of bibliographic ontologies made by computer scientists alone.

References


A Conceptual model for an OWL ontology to represent the knowledge of transmedia storytelling

Abstract
This paper proposes a conceptual model and an OWL ontology for the representation and knowledge organization of transmedia narratives and the creation of RDF datasets. The authors have adopted an approach based on the development of a flexible conceptual model for the management and representation of information accessible in a network environment. The conceptual model identifies a series of entities, attributes and relationships to describe, organize and interrelate the knowledge of transmedia contents. From the conceptual model an OWL ontology has been developed in which SKOS has been widely used in the ontology to separate the conceptual description of resources. The conceptual model enables the design of architectures for the consumption of contents and the ontology offers a first conceptual level for organizing knowledge of transmedia narratives.

1. Introduction
Digital platforms for audiovisual content are changing the concept of such products. Content related to movies and television series posted on blogs, wikis and social networks reflect a high level of reader participation. Consequently, users expand the narrative environment by developing content about characters, events, places, etc. (Warren, Wakefield and Mills 2013, 69).

Audiovisual content can be based on literary works, comics, video games. They also provide a source for extending narrative structures, creating authentic fictional micro-universes with a high degree of complexity and plot coherence (Long 2007, 21).

Transmedia storytelling uses multiple platforms to develop new narrative resources, with external contributions, adapted to the nature of the medium in which they are published (Jenkins 2003, 2011). Transmedia storytelling also includes theatrical performances, meetings, exhibitions and other types of social and artistic live events (Edmond 2014) thus becoming authentic cultural phenomena that transcend the original content itself.

Audience participation is essential to create transmedia storytelling and resources associated with content (Deuze, 2006; Scolari, 2009) creating a dynamic in which the user becomes a prosumer, that is, simultaneously producer and consumer of content (Fernández 2014).

The interrelation of content is defined from descriptive metadata. The capacity of systems for the exchange, description and retrieval of information is based on the level of syntactic, structural and semantic interoperability (Veltman 2001, 161). The Semantic Web offers a multilevel architecture for a high semantic expressiveness of
data. Therefore, various technologies are used: the RDF data model and serialization formats, RDF Schema, OWL, SPARQL, etc.

This paper defines a conceptual model to formalize and describe transmedia storytelling and an OWL ontology for its representation as linked open data datasets. Through this proposal it would be possible to exchange this type of information between systems, create comparative studies of transmedia storytelling or publish datasets for consumption by web services or multiplatform applications.

2. Proposal for a conceptual model of transmedia storytelling

The use of tools to represent and formalize scripts is an idea explored in the editing processes of some cinematographic productions. Examples such as those of *The Matrix* and *Prometheus* are paradigmatic in this regard (Kallay 2013, 86-95). The processing and exploitation of this information in digital environments requires structured data, based on conceptual models, to represent abstractions of ideas or objects from the real world (Solodovnik 2011, 5). An example of these models in the audiovisual field are EBUCore / EBU-CCDM (EBU, 2016, 2017).

Some authors have analyzed the different elements associated with the conceptualization of contents elaborated by transmedia storytelling (Rampazzo, 2013). Most of these studies establish that an audiovisual resource in digital format can be described under a multilevel approach. This allows the definition of different scenarios to reuse descriptive metadata associated with different levels of granularity.

The proposed model describes, at a basic level, resources, certain narrative elements (places, characters, actors, periods, events) and different types of transmedia storytelling relationships between works and audiovisual content. Entities, attributes and relationships are identified to describe and organize content resources. The aspects regarding the production, distribution or commercialization of content remain outside the scope of this proposal.

2.1. Resource

The entity "Resource" represents any type of object, such as a “Creative Work” or an “Event”. In this context, a creative work is understood as a film, television series, book, comic, website, document, etc. Events include any type of activity, such as theatrical performances, exhibitions, festivals, etc. The resources can be published by an editor (“Publisher”) and classified by category ("Category"). It is possible to define semantic relationships between categories. Resources can also be associated with a specific context, such as a franchise, brand or narrative universe ("Context"). The relationship "has related resource" represents associative relationships between resources, while "has part" is used for part-whole relationships. Creative works and events are linked to each other with the relationship "has event". The model allows the definition and organization of different types of creative works and events by assigning
The organization of the TV series is represented by the sub-entities "Series", "Season" and "Episode". A series is linked to the first season through the relationship "has first season". The sequence between seasons is defined with the relation "next season". In a similar way, the relations between seasons and episodes are defined using "has first episode" and "has next episode" respectively.

A key aspect is the narrative relationships between resources that define transmedia
methods to expand the story content universe. The conceptual model represents this characteristic through “storytelling” relations. Based on the work of Jung and O-Joun (2017), this relationship can be specialized in more specific ones.

The sequential organization of the creation, publication and development of narrative timelines for content resources is defined by the relations "have next in creation", "have next in publication" and "have next in timeline story", respectively.

The model defines the sub-entity "Fragment" to establish divisions within a work. It is also possible to refer to a certain (usually significant) event that takes place during the narration of content.

2.2. Users

Users are an essential part of the conceptual model, since transmedia storytelling requires user interaction as a prosumer. The "User" entity identifies individual users that can be organized into groups or communities ("User Group"). The entity "Participation" links users with resources in various ways. The type of participation is defined using the "have participation type" relationship and the entity "Participation Type". This mechanism is able to indicate when a user makes comments, evaluations, suggestions, collaborations or when he is spectator of an event or consumes content.

Figure 2: Diagram D-2 representing entities and the interaction of users with resources
Source: own
The model (see Figure 2) contemplates that the user defines routes or itineraries (entity "Itinerary") to consume resources. Through this feature, the user could make a selection of resources to start exploring a narrative universe and establish multiple paths of visualization or consumption of resources. It would also be possible to define itineraries pre-established by the publishers so that the user only selects the one in which they are most interested according to the type of resource, length, etc. A possible application of this feature can be found in some TV series with a main narrative line that develops in certain chapters and others that are outside of that line.

2.3. Agents, cast and contributions

In the proposed model, an agent is any person or organization, real or fictitious, that forms part of the argument or participates in its creation (see Figure 3). The conceptual model accordingly distinguishes four types of agents: persons (entity "Person"), groups or organizations (entity "Group"), characters (entity "Character") and factions, organizations or fictitious groups (entity "FictionalGroup").

Figure 3: Diagram D-3 representing agents, distribution and contributions
Source: own
Individual agents ("Person" and "Character" entities) can be grouped using the relationship "has member". In turn, hierarchies can be defined among group agents using the relationship "has subgroup".

Agents have a direct relationship with content resources (property "related agent") or through a relationship depending on whether they are part of the cast or a team of creators and collaborators. This is the function of the entities "Contribution" and "Cast". While the former defines the contributions of people based on the function (role) they have played, the latter establishes the cast, indicating characters and the actors playing them. In addition, it is possible to define relationships between agents. This type of relationship can be used to define kinship or professional ties, etc. between different agents.

3. OWL ontology for the publication of transmedia storytelling datasets

Ontologies provide a logical-conceptual model for a knowledge domain, defining, at different levels of formalization, the meaning of the classes of objects, the relationships between them and the properties that describe them.

There is an intensive use of ontologies in the context of the semantic web, and some of them, such as SKOS (Miles and Bechhofer 2009) or the Dbpedia ontology (Lehmann et al. 2015) are fundamental elements in the development of applications and linked open data datasets. Ontologies also follow the principles of interoperability and reuse, which underlie the logical aspects of resource description and the interrelationships between them.

The authors have developed an OWL ontology from the conceptual model, whose namespace (also valid for downloading) is http://purl.org/umu/tso/ with documentation available at http://skos.um.es/TR/tso/.

SKOS has been used to represent vocabulary entities and controlled terms (categories, roles, types of events and works, types of user participation). The use of SKOS thesaurus or classifications for the definition of typologies provides the ontology with greater flexibility and efficiency of use.

Figure 4 shows an RDF graph with an example of the use of an ontology. The corresponding prefixes have been used to simplify their reading. The prefix "tso" has been used to represent the elements of the proposed ontology. The prefix "ex" has been used to represent resources in a fictitious dataset.

The classes of the resources are indicated in parentheses under the corresponding IRI. The book "The Fellowship of the Ring" by J.R.R. Tolkien has been defined as part of "The Lord of the Rings" using the property tso:hasPart. Peter Jackson's film of the same title is defined as a version of that book with the prequel (tso:hasPrecuel) "The Hobbit: An Unexpected Journey". Two fragments (tso: Fragment) have been defined, both in the novel and in the film "The Fellowship of the Ring". In the first case, the
fragment refers to book II, chapter 5 of the novel. The relationship between the movie and its corresponding fragment is defined as an event using the tso:hasEvent property. Its initial and final position is indicated with the properties tso:startPosition and tso:endPosition, respectively. To define a relation between both fragments, the property tso:hasRelatedResource is used. The characters of Gandalf and Balrog are related to the fragment of the book by tso:hasRelatedAgent. The resource ex:castGandalf allows a link to be established between Gandalf and Ian McKellen, the actor who plays the character, and relate that performance to the movie.

Figure 4: Graph RDF with an example of application of the ontology
Source: own
4. Conclusions: towards an architecture for consumption of transmedia data and content

In the context of transmedia storytelling, the advantages of structured data are centered on its reuse in a multiplatform context. The exploitation and enrichment of these datasets offer possibilities of extending information beyond the original data. Therefore, the combination of audiovisual digital content platforms and external datasets based on RDF (Urakawa et al. 2016) is of great interest.

The development of semantic platforms is a field that has been under development for years (Speicher, Arwe and Malhotra 2015). However, the incorporation of transmedia content can offer new tools that integrate streaming video services with added value information during the reproduction of contents. Standardization of such information would be essential between producers and consumer services.

Interconnections are defined and formalized by an architecture wherein transmedia content and data are consumed. Within this architecture, narratives are created from audiovisual contents, users, events and objects from the socio-cultural heritage of the real world. The digital context allows the deployment of content access and consumption platforms and also allows the publication of structured, interoperable and, preferably, open data. This data can lead to all kinds of content, applications and services on the web and, ultimately, could give access to streaming content. A fundamental element is the structured data warehouses that would be connected to such content. It could also be published as linked open data made available by third parties for the development of applications. In the same way, the web is a source for the enrichment of metadata, from the definition of links to web resources or with other RDF datasets, such as DBpedia, Wikidata or Geonames.

The proposed conceptual model and OWL ontology define a first level in the description and interoperable representation of transmedia storytelling. In this level, the description of the basic elements of the resources are the basis for much more detailed descriptions. As a line of future research, a second level could be developed to represent the most elaborated plot aspects that would enable the design and development of plots, characters, complex user interactions, etc.

There are many tools used to design transmedia storytelling such as Twine, Mapstory, Zeega, Klynt, Moveable Feast, If This Than That, etc. The exchange of interoperable data between different software would allow the creation and application of semantic standards from the first moment of the creative process.

In the near future, the integration of content and data structured by semantic technologies could offer more efficient and usable platforms and services. The transparent and integral application of systems would result in a better user experience, which is a key condition for the success of any technology.
References


Connecting KOSs and the LOD Cloud

Abstract
This paper describes a specific project, the current situation leading to it, its project design and first results. In particular, we will examine the terminology employed in the Linked Open Data cloud and compare this to the terminology employed in both the Universal Decimal Classification and the Basic Concepts Classification. We will explore whether these classifications can encourage greater consistency in LOD terminology. We thus hope to link the largely distinct scholarly literatures that address LOD and KOSs.

1. Introduction and motivation
Our research\(^1\) involves comparing the terminology employed within the Linked Open Data (LOD) Cloud with terminology employed within two KOSs: the Universal Decimal Classification (UDC) and the Basic Concepts Classification (BCC). In doing so we will connect two quite distinct literatures and communities of practice: the Semantic Web (SW) community, which has tended to be centered in computer science, and the knowledge organization (KO) community. In the SW community there have been increasing efforts to curate and preserve the machine-readable knowledge items as published on the Web using linked data formats (Beek, Rietveld et al. 2014; Beek et al. 2014). Controlled vocabularies play a prominent role in these efforts. They provide a way to index the knowledge graph, and they represent a semantically enriched layer in this graph. In knowledge organization (KO), systematic studies of KOSs have been proposed already (Tennis 2012), and such studies have also been executed for a number of small samples.

The promise of the web-based LOD Cloud is to free up data, metadata and information to a large extent from what often is called “data silos” – isolated information systems, which come with their own domain-specific knowledge organization systems, and are often barely interoperable. The LOD Cloud promises to deliver machine-readable KOSs and their implementation in a way that enables easy cross-linking. For example, the platform GeoNames (http://www.geonames.org) publishes about eleven billion place names in machine readable form, and has been used by many other services to relate a term like “New York” to a specific geographic reference, which in turn enables other services to link other names to this location, e.g., “City of New York,” “New York City,” or the historic term “Nieuw Amsterdam”.

To be able to compare the different terminologies expressed in vocabularies, one

\(^1\) Digging Into the Knowledge Graph, 2016 Digging Into Data Challenge. Available at: https://diggingintodata.org/awards/2016/project/digging-knowledge-graph.
first has to have an overview of them. Hence, our research involves the initial step of
surveying the terminologies that are currently employed in linked open data. This will
result in an atlas of vocabularies.

The SW holds the promise that different information repositories can all be encoded
in the flexible graph-based representation language Resource Description Framework
(RDF). Atomic statements in RDF take the form of triples, which are composed of a
subject, predicate, and object term. RDF relies on URIs in order to assign universally
unique names to concepts and instances. Since RDF names also denote locations on the
internet, it uses RUIs for both naming and locating. By navigating to a URI location,
software agents are able to extract the description of the entity denoted by that URI.
Besides these syntactic and infrastructural properties, RDF also has a model-theoretic
semantics that allows inferences to be drawn mechanically across different sets of
information. If, for example, one website contains the RDF triple “Birds have wings”
and another website contains the RDF triple “Penguins are birds,” a computer can infer
that “Penguins have wings.” But this will only work if the same, or interoperable,
terminology is employed. At present a wide variety of controlled vocabularies are
employed across the LOD Cloud, but their formal semantics, including the inferences
that follow from it, are not yet studied on a large web-scale.

Areas that make extensive use of SW technology include the humanities and the arts,
as well as the social sciences (e.g., Hyvönen 2012). Mirroring the large variety of social
and cultural phenomena in these fields, we find very specific, context-rich vocabularies
developed by research communities as well as curators of collections. Increasingly,
traditional curators of such vocabularies (e.g., The Getty Research Institute) provide
their vocabularies as LOD. In consequence, big data projects in the social sciences and
humanities embrace SW technology (e.g., Hyvönen 2012). The ultimate goal of the
collaboration in this project is to enhance the findability of facts and vocabulary used
in the LOD Cloud and to enable scholars in the social sciences and humanities to find
the right points to connect to when publishing LOD.

In the standardization of the SW (Horrocks 2003), a conscious choice was made to
not create a specific upper ontology. Instead, SW standards like the Web Ontology
Language (OWL) define a logical language that is devoid of content. Specific
ontologies can be defined in terms of the logical primitives provided by these
languages, but no specific conceptualization is prioritized. Specifically, the idea is that
alternative, complementary, and even contradictory ontologies can be defined. This
implements one of the core goals of the SW, which is to ensure that “anyone can say
anything about anything”.

---

In addition to the ability to encode multiple ontologies, the SW was designed in order to scale to facilitate the world-wide interchange of knowledge. Taking ideas from the World Wide Web (WWW), which is very successful in facilitating the world-wide interchange of documents, scalability is believed to be hampered by the existence of a centralized authority that coordinates the creation, distribution and retrieval of specific ontologies.

An inevitable result of these design goals is that there is currently not a complete overview of the ontologies that are published on the SW, although there are manually curated collections like Linked Open Vocabularies (LOV) (VanDenBussche 2017). Even though there is a limited overview of which ontologies are present on the SW, datasets that are part of the LOD Cloud are known to often reuse a core set of popular existing vocabularies (Schmachtenberg 2014). In order to obtain a better overview of the ontologies and datasets that are currently published as LOD, the LOD Laundromat was developed at VU University Amsterdam.

But the collection of web-based vocabularies is only a first step. We will then proceed to compare the terminologies that are published in the LOD Cloud with the controlled vocabularies of the UDC and BCC. Note here that the challenge of interoperability across the LOD Cloud is itself a KO challenge; there is no explicit coordination between LD vocabularies.

There has been limited communication between the KO community and those active in developing the Semantic Web. We chose the UDC and BCC because these classifications have explicitly grappled with interdisciplinarity, and have pursued a faceted approach to classification (on BCC see Szostak 2013). The potential of the Semantic Web will best be realized if connections can be drawn across repositories. We thus wonder whether KOSs that strive to facilitate interdisciplinarity can play a key role in encouraging interoperability in the LOD cloud. Can the terminology employed in the LOD cloud be connected to KOS controlled vocabularies? Can the hierarchies and other relationships recognized within KOSs be used to structure terminology in the LOD cloud?

We will use those two generic classifications (UDC, BCC) as reference systems to develop generic principles of indexing. We will use high-level topical categorisations (similar to the UDC classes) and facets (similar to UDC auxiliaries such as place, time, peoples, forms, languages, etc.). We will contrast this with the phenomenon-based approach of the BCC, and ask questions of What (is studied)?, Why, Who, Where, and When? These categorisations will be tested in the archived version of the LOD Laundromat and eventually implemented in the open web services of the “living” LOD Laundromat. In particular, we will explore how general classification systems such as the UDC and BCC can be used to index linked data in a way that allows searching for concepts across domains, without becoming lost in the richness of the KOSs embedded
in the LOD. In other words, we aim at a kind of union catalog for the LOD Laundromat snapshot, which will also be archived along with the LOD Laundromat data collection itself. One key question we hope to investigate is how interdisciplinarity is present and expressed or hidden and undiscovered.

At present, anyone wishing to code data for the SW has to choose among a bewildering array of sources of terminology. The choices made will determine to which other data repositories a computer can connect your data. Our research can potentially ease the choice facing those wishing to employ LOD and expand the degree of interoperability. We hope, in particular, to develop recommendations for LOD publication for communities in the social sciences and humanities (SSH), with emphasis on improving the re-use of existing vocabularies (among which we will encourage interoperability). We will identify, evaluate and index SSH-relevant vocabularies by mapping clusters of similar meaning onto KOSs.

Though implications for the SW are perhaps most obvious, our research also has important implications for KO. If KOSs can play a critical role in encouraging interoperability across the SW, then the field of KO gains an important new audience for its work. Note that the premise of the SW is that data of all types need to be explicitly encoded in terms of formal languages such as RDF and OWL. In other words, the SW is grounded in the recognition that there are limits to what can be discovered by simply searching un- or weakly-structured natural language texts. The KO community’s longstanding efforts to develop structured controlled vocabulary at times seems to be overshadowed by search algorithms that search natural language texts rather than structured metadata, but the SW potentially places KO at the center of future developments in search engines. In addition, research has shown that classifications themselves form navigable knowledge networks among the resources to which they are linked (Suchecki et al. 2012; Smiraglia et al. 2013).

Much effort is undertaken in the KO domain to bring KOSs into use in the LOD Cloud (e.g., Baca and Gill 2015). As mentioned above, there is effort to link general controlled vocabularies, such as The Getty Vocabularies, to the LOD Cloud. There are definite advantages in vocabulary mapping for people-centered properties (what librarians call “authority control” of names), for LOD, to alleviate the problems of property proliferation in LOD environments. The discourse concerning the SW reveals a research agenda for KOSs including direct linkage of domain-centric ontologies within the LOD Cloud, including most importantly for this project, vocabulary alignment. We hope to provide advice on how KOSs might be revised to reflect and serve the LOD Cloud (especially from the perspective of interdisciplinarity).

It should be stressed that the proposed research will provide a much-needed link between LOD and KOSs. By mapping one onto the other we can compare the structure of the two. KOSs always employ some sort of logical structure with enumerative
capability: the idea that a place must be found in the KOS for all works or ideas. In addition, the traditional use of “literary warrant” allows a KOS to point directly to the documentary source of a particular concept. Comparing LOD clusters with a KOS can indicate where a particular KOS needs to be amended. That is, LOD clusters provide literary warrant for extending enumeration and clarifying the KOS. In turn, the mapping can suggest how LOD can be better structured/indexed to facilitate the practice of actually linking data. We can thus harness the wisdom of the KO community to the important practice of achieving interoperability or even consensus on LOD terminology. To achieve this interplay with respect to the BCC we must render the BCC into LOD terminology and then compare the result with the clusters of LOD terminology we obtain.

The next three short segments of this paper describe deliverables from our research as of January 2018, midway through the first year of a three year project.

2. Toward a KOS observatory

A challenge frequently articulated across the KO and SW communities is the ability to track and maintain access to changing KOSs across time and across applications. In addition to Beek et al. (2014), Tennis (2002; 2007; 2012; 2015; 2016) has been the most prominent catalyst for the KO community. Two KNOWeSCAPE workshops were held in Amsterdam (2015) and Malta (2017) to bring together experts from the KO, SW, publishing and digital humanities communities to prioritize objectives for visioning and creating an observatory for KOSs. Following on these workshops and in conjunction with our research, an effort to map a small initial set of KOSs has been undertaken by the DANS (Data Archiving and Networked Services) division of the Royal Netherlands Academy of the Arts and Sciences (KNAW). Our team is working on a project of larger scope than the similar Basel Register of Thesauri, Ontologies & Classifications (BARTOC, https://bartoc.org), but consulting BARTOC along with standard bibliographic and internet resources we have created a template and begun building a database of KOSs. Figures 1a-b show parts of the experimental template for *Art & Architecture Thesaurus*.

---

Figure 1a: Observatory template for Art & Architecture Thesaurus, part 1, edited

<table>
<thead>
<tr>
<th>Creator(s)/Curator(s)</th>
<th>Maintenance organization</th>
<th>Format(s)</th>
<th>Physical Location</th>
<th>Online Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toni Petersen</td>
<td>Getty Art History Information Program</td>
<td>Printed Book.</td>
<td>KOS.2!A1:A13</td>
<td>n/a</td>
</tr>
<tr>
<td>Toni Petersen</td>
<td>Getty Art History Information Program</td>
<td>Printed Book.</td>
<td>KOS.2!A26:A27</td>
<td>n/a</td>
</tr>
<tr>
<td>Toni Petersen</td>
<td>Getty Art History Information Program</td>
<td>Digital, 6 computer discs; 3 1/2-5 1/4 in. + 1 Introduction to the art and architecture thesaurus, 2nd ed. (250 pages ; 27 cm) + 1 quick-reference card + 1 user's manual for the Authority references tool ([126] pages ; 27 cm)</td>
<td>KOS.2!A26:A27</td>
<td>n/a</td>
</tr>
<tr>
<td>Toni Petersen</td>
<td>Getty Art History Information Program</td>
<td>Digital, 3 computer discs; 3 1/2 in. + 2 user manuals (27 cm) + 1 quick reference card (27 x 57 cm folded to 27 x 19 cm) + 1 general information card (24 x 57 cm folded to 24 x 15 cm) + 1 demo disc</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Figure 1b: Observatory template for Art & Architecture Thesaurus, part 2, versioning data, edited

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Schema Name/Title</th>
<th>Earlier versions (editions) …</th>
<th>History of versioning:</th>
</tr>
</thead>
<tbody>
<tr>
<td>KOS.2</td>
<td>Art &amp; Architecture Thesaurus; AAT</td>
<td>n/a</td>
<td>1st Edition</td>
</tr>
<tr>
<td>KOS.2.1</td>
<td>Art &amp; Architecture Thesaurus; AAT</td>
<td>1st edition 1990</td>
<td>1st Edition: Supplement 1</td>
</tr>
<tr>
<td>KOS.2.2</td>
<td>Art &amp; Architecture Thesaurus; AAT</td>
<td>1st edition supplement 1, 1992</td>
<td>2nd Edition</td>
</tr>
<tr>
<td>KOS.2.5</td>
<td>Art &amp; Architecture Thesaurus; AAT</td>
<td>2nd edition 1994</td>
<td>Online</td>
</tr>
<tr>
<td>KOS.2.7</td>
<td>AAT-Deutsch</td>
<td>Online</td>
<td>Online, Version DE</td>
</tr>
<tr>
<td>KOS.2.8</td>
<td>TAA; El Tesuario de Arte &amp; Arquitectura</td>
<td>Online</td>
<td>Online, Version ES</td>
</tr>
</tbody>
</table>
Because of space limitations these illustrations have been edited. By experimental template, we mean that during the course of collecting information about KOSs used in the social sciences and humanities domain, our own KO–as embodied in the template is changing, depending on the information available on the KOSs, and our growing insights.

However, it is clear to see that because of our aim to link these KOSs to the LOD Cloud as well as to enable version tracking we are collecting detailed data on location and versions for each system. It is our plan to archive our own data in the DANS EASY online archiving system for research data (https://easy.dans.knaw.nl/ui/home).

3. Classification as LOD

A major goal of the first year of research is to create both generic classifications (UDC and BCC) as linked data. This research is progressing in two separate streams. BCC is being rendered as LOD by the University of Alberta project team. The UDC is a proprietary system, however, so mixed solutions will be required. The UDC Summaries online, which are publicly available via the internet already exist as LOD (http://udcdata.info/). It is important that the entire UDC be rendered as linked data, but the proprietary portions may remain behind paywalls for use by licensed subscribers.

4. Classifying concepts (not documents)

Classifying concepts is different from classifying documents. Typical document classification is tied to a summary of the overall subject of a document, in order to place the document among related texts. Our goal, however, is to use both classifications to point to specific concepts, rather in the style of the famous grinder metaphor of Paul Otlet (see Figure 1 in Smiraglia and van den Heuvel 2013, 363). In other words, we hope to create networked linkages in the LOD cloud among concepts, their classification nodes, and the SW resources to which each are linked. Research comparing the use of the two classifications empirically in a set of OCLC WorldCat data has revealed the problem of concepts that are less explicit in UDC because they are hidden hierarchically (Szostak and Smiraglia 2018). The phenomenon-based BCC is more efficacious for directly pointing to concepts. The two classifications together, however, provide the best of both worlds by pointing directly to concepts and also placing them in disciplinary contexts.

5. Conclusion

The encounter between SW and KO comes naturally with the need to sort out terminology, concepts, and epistemic values. One task of this project, for which this paper is an example, is to provide “translations” from one knowledge domain into the other, to create a commonly shared understanding and enabling communication; in short, to establish a trading zone as science and technology studies would call this
Another precondition to make specific KOSs operable in the LOD cloud is their publication as LOD. This task can rely on some traditions of the adoption of LD principles for those curating KO. Much more challenging is the task to explore to which extent those new generic KOSs can be made of use to a) increase findability of KOSs in the LOD; and, b) to connect UDC and BCC to other parts of the LOD so that their power of expressing concepts and describing phenomena in a manner that is universal, cross-cultural and cross-domain can be turned into a navigation tool inside of the LOD.

Acknowledgment

We would like to acknowledge Gerard Coen, DANS, for contributions to the initial mapping of KOSs including those shown here in Figures 1a-b.

References


Richard P. Smiraglia, Rick Szostak

Converting UDC to BCC: comparative approaches to interdisciplinarity

Abstract
The knowledge organization domain has been turning its attention increasingly to problems of interdisciplinarity. Recently we have attempted to explore the approaches to interdisciplinarity represented by the synthetic and faceted Universal Decimal Classification (UDC) and that of the phenomenon-based Basic Concepts Classification (BCC). The questions for research are: how do both classifications express the same sets of concepts, what are the specific advantages or disadvantages of disciplinary versus phenomenon-based classification in the gathering of concepts, and how can these classifications be used to generate interdisciplinary ontologies for the Semantic Web? The study reported here takes an empirical approach to the comparison of the UDC and the BCC assigned to a set of documents found in the OCLC WorldCat. The present study demonstrates both the greater economy and greater conceptual precision in the phenomenon-based BCC. The network analysis suggests that there is great navigational strength in both approaches.

1. Interdisciplinarity, phenomena, concepts and KOSs
The knowledge organization (KO) domain has been turning its attention increasingly to problems of interdisciplinarity, both in the sense that the domain is working to create knowledge organization systems (KOSs) that bridge disciplinary boundaries, and in the sense that the domain has begun to seek truly interdisciplinary knowledge organization solutions. One powerful approach to interdisciplinarity increasingly demonstrated is the structure of classifications around individual phenomena rather than in disciplinary groupings. The obvious advantage of phenomenon-based systems for interdisciplinarity is gathering by phenomenon despite disciplinary or epistemic stance.

Recently we have attempted to explore the approaches to interdisciplinarity represented by the synthetic and faceted Universal Decimal Classification (UDC) and that of the phenomenon-based Basic Concepts Classification. Both are large-scale general classifications capable of expressing complex concepts with precision and subtlety. The questions for research are: how do both classifications express the same sets of concepts, what are the specific advantages or disadvantages of disciplinary versus phenomenon-based classification in the gathering of concepts, and how can these classifications be used to generate interdisciplinary ontologies for the Semantic Web? The first two questions have been examined briefly in an exploratory study that compared classification strings assigned by both classifications to the same set of documents (Szostak and Smiraglia 2017) and mapped the network relationships among facets in both sets of strings (Smiraglia and Szostak 2017). The third question is one trajectory of a large-scale research project connecting KOSs and the LOD Cloud.
The study reported here takes an empirical approach to the comparison of the UDC and the BCC assigned to a set of documents found in the OCLC WorldCat.

We do not need to rehearse the origins of the Universal Decimal Classification here (see Szostak and Smiraglia 2017), except to recall that its founder, Paul Otlet, had as his goal the specific ordering of concepts extracted from deconstructed texts. Otlet generated, then, a classification of knowledge that is commonly used to assign many UDC strings to any particular document, in order to precisely identify topical phenomena. Empirical research has demonstrated the power of this approach to classification using the UDC (Smiraglia 2016; Scharnhorst et al. 2016), one aspect of which is its generation of a network linking phenomena within the classified set of documents represented by UDC strings.

The Basic Concepts Classification (BCC) was created by Rick Szostak with the explicit goal of providing a means to classify documents (and objects and ideas) with respect to the phenomena they study. The BCC has evolved through the growth of schedules of (mostly verb-like) relators and adjectival/adverbial properties added to the original schedule of phenomena. Documents (objects, ideas, concepts) can be classified with combinations of phenomena, relators and properties. Szostak (2016; 2017) suggests subject classifications should follow basic grammatical structures in combining these three types of term; such subject classifications will thus appeal to the linguistic facility of both classifiers and users.

2. Methodology

We compiled a set of documents classified using the UDC taken from a random sample drawn from the OCLC WorldCat. Nine million UDC strings from the WorldCat, representing essentially a dump of the entire population of UDC in the WorldCat at one point in time, were originally downloaded by the OCLC Office of Research (Scharnhorst et al. 2016). These are representative of classification provided by mostly European UDC libraries using the WorldCat and assigning UDC to mostly scientific and technical late twentieth century works. Through pilot studies and analysis of earlier empirical studies it was determined that a sample drawn at random would need 381 cases to provide results generalizable at 95% confidence ±5%. We deciphered each

---

1 Digging Into the Knowledge Graph, 2016 Digging Into Data Challenge. Available at: https://diggingintodata.org/awards/2016/project/digging-knowledge-graph/


3 This sample size formula based on a proportion with a known population size is: 
   \[ n = \frac{z^2Np(1-p)}{NE^2 + z^2p(1-p)} \]
   where 
   \[ n = \text{sample size}; z = \text{curve value for the confidence interval (.95)}; N = \text{items in the sampling} \]
UDC string, deconstructing each element of it and citing an example from the WorldCat of a resource to which the particular string had been assigned. We then provided BCC classification to match each UDC string. In Szostak and Smiraglia (2017) we included cases from Portuguese sources (see Scharnhorst et al. 2016) to demonstrate the ability of BCC to capture the entire content and context in a single string. For each case one BCC string was assigned, but the range of number of UDC strings was 1–7; the ratio was 1 to a mean of 2.2, for a multiplier of .45. In the present study we are comparing UDC strings drawn from the WorldCat rather than classification assigned to particular resources. Thus, the data reported here do not further demonstrate this aspect of the relationship between the expressivity of the two classifications. Our emphasis here is on conceptual expression and network relationships.

3. Results

We were able to decipher 382 UDC strings using UDC Online (http://udc-hub.com/) and Attila Piri’s UDC-parser (http://piros.udc-interpreter.hu/#). Thus our data yielded results that describe well the UDC-classified content of the WorldCat. We recorded data on ontogenetic questions to be used in a future paper. Our analysis here visits a) the population of both UDC and BCC concepts; b) string length and number of terms; c) comparative conceptual precision; and d) network analysis.

3.1. Population of the UDC and BCC

One informative metric about a classification can be the degree to which its classes are populated in any given classified environment. Classifications like the UDC are considered general classifications to the extent that they are thought to be hospitable to all major disciplines of knowledge. By visualizing the population of the UDC main classes in our study we create a contextual picture of the degree to which various disciplines are heavily used or, alternatively, little used. Figure 1 shows the population of the UDC main classes and two-digit disciplines derived from the sample used in this study.

\[ p = \text{expected proportion (.45) ratio of BCC string size to UDC string size in Szostak and Smiraglia (2017a); } E = \text{tolerable error (.05).} \]
The distribution is consistent with earlier visualizations of the population of the main classes in the 2009 WorldCat UDC dump as summarized in Scharnhorst et al. (2016) and Smiraglia (2016), and this consistency helps to demonstrate the generalizability of the data in our sample. The largest disciplinary clusters and hence the majority of the works (approximately 30% each) fall in the social sciences and the applied sciences, and another 30% comprises natural sciences, literature and the arts. There were sixty-four disciplinary combinations of two digits in the sample. From this visualization we can see that the social sciences are predominantly populated by education, economics and law, that the applied sciences comprise medicine, engineering, communication and agriculture, and the other large disciplines represented are mathematics and literature.

The BCC is phenomenon-based rather than discipline-based, so that analysis of the first character of each string would point to classes of phenomena. Eighteen characters were found in first position. The list, shown in Table 1, is remarkably different from the basic disciplinary classes we saw in UDC.

<table>
<thead>
<tr>
<th>T. Technology and science</th>
<th>110</th>
<th>N. Non-human environment</th>
<th>13</th>
<th>F. Flora and fauna</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>→. Relation, Influence on</td>
<td>44</td>
<td>H. Health and population</td>
<td>10</td>
<td>G. Genetic predisposition</td>
<td>3</td>
</tr>
<tr>
<td>P. Politics</td>
<td>44</td>
<td>S. Social structure</td>
<td>9</td>
<td>′. Relation, Of</td>
<td>2</td>
</tr>
<tr>
<td>E. Economy</td>
<td>35</td>
<td>I. Individual differences</td>
<td>7</td>
<td>↑. Relation, Increase</td>
<td>1</td>
</tr>
<tr>
<td>A. Art</td>
<td>33</td>
<td>Σ. Relation, Collection of</td>
<td>6</td>
<td>M. Molecules and elements</td>
<td>1</td>
</tr>
<tr>
<td>C. Culture</td>
<td>29</td>
<td>‼. Relation, Of type</td>
<td>5</td>
<td>X. Mathematical concepts</td>
<td>1</td>
</tr>
</tbody>
</table>
Population of these classes is shown proportionally in Figure 2 for comparison with the population of the UDC classes. (Note: One reason for the large number of entries beginning with T is that different types of document, such as dictionary or textbook, are captured within class T.) This visualization gives us a remarkably different, and perhaps fuller picture of the conceptual content of the classified collection. The largest cluster of 31% for Technology and science mirrors the largest cluster in the UDC. Politics, economy, art, culture and non-human environments are all clearly articulated in clusters ranging from 4% to 12% of the total. But causal relators, which are always affixed as auxiliaries in UDC lead in many cases in BCC, with “influence on” constituting the second largest cluster equal in size to that for politics. The vague disciplinary clustering is replaced with a more vivid description of the classified collection.

3.2. UDC string length and number of terms

The mean UDC string length is 7.7 characters and 1.5 terms; the median is 6 characters and 1 term; the mode is 3 characters and 1 term. Sorted by combination of string length and number of terms we discovered there were 53 different combinations. One-character strings (e.g., “3” Social sciences) occur only 3 times and tend to be assigned to nonbook materials. Three-character single-term strings occur 59 times (17%), which is the largest frequency of occurrence. Five-character one-term strings occur 42 times, the second highest frequency (12%). Longer strings having 14-22 characters always have two or more terms. Strings with 23 characters always have three or more terms, and the longest strings (more than 24 characters) always have four or more terms. Strings combining two terms are fairly common; more complex combinations are rare.
3.3. BCC string length and number of terms

The mean BCC string length is 10.5 characters; the median is 9 and the mode is 7 characters. The mean number of terms per string is 3.5; the median and mode are 3; the number of terms ranges from 1 to 9. (Where the classification indicates “Cutter number” we have counted it as one term with two characters, e.g., an alphanumeric “B3.”) There is greater disparity among string lengths; the largest occurring frequency is 7 characters, which occurs 42 times (20%); the second most frequently occurring is 12 characters, which occurs 34 times (16%). The range in character length is from 2 to 29 characters. The length hovers around the mean; extremes (25 or more characters, but oddly also 2 or 5 characters) are rare. The most frequently occurring combination is 7 characters and 3 terms (7%), followed by 3 characters and 1 term (5%). Thirty-five strings (36%) ranging from 3 to 29 characters in length, have only 1 term, which is a further reflection of the specificity of the phenomenon-based classification.

3.4. Comparative conceptual precision in UDC and BCC

Analysis of the text of the deconstructed strings reveals several large conceptual clusters in the data. For the purpose of this paper we isolated two, “military affairs” and “economic outputs” in order to compare the classified strings. Table 2 shows the comparative results.

<table>
<thead>
<tr>
<th>Table 2: Concept clusters “Military Affairs” and “Economic Outputs”</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Military Affairs</th>
<th>Economic Outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>UN peacekeeping doctrine</td>
<td>marine pollution and sea life</td>
</tr>
<tr>
<td>355.4</td>
<td>591.9(26)</td>
</tr>
<tr>
<td>3 Social science—35 Public administration—355</td>
<td>5 Natural sciences—59 Zoology—591 General zoology—591.9 Geographic distribution of animals—(26) common auxiliary of place 2 Physiographic designation—26 Oceans, seas and interconnections (EO→ga QN) → F&gt;NT3oa (Pollution = economic output does not facilitate the quality of the environment)(affects)(life)(in)(oceans)</td>
</tr>
<tr>
<td>Military affairs—355.4 War operations TF(P15a)→gerx</td>
<td></td>
</tr>
<tr>
<td>TF professional field—TF(P15a) military field—</td>
<td></td>
</tr>
<tr>
<td>^→gerx associated with ending conflict</td>
<td></td>
</tr>
<tr>
<td>Atlantic convoys</td>
<td>Mary Cassatt prints</td>
</tr>
<tr>
<td>355/359</td>
<td>76</td>
</tr>
<tr>
<td>3 Social science—35 Public administration—355</td>
<td>7 Arts—76 Graphic art, printmaking</td>
</tr>
<tr>
<td>Military affairs—359 consecutive extension to Armed</td>
<td>EO9821215 ^ A</td>
</tr>
<tr>
<td>forces</td>
<td>(economic output of prints)(associated with)(art)</td>
</tr>
<tr>
<td>TF(P15a)∑EO925111^PI5n&gt;NT3oa</td>
<td></td>
</tr>
<tr>
<td>TF(P15a) Military science—∑EO925111 &gt; NT3oa</td>
<td></td>
</tr>
<tr>
<td>Convoy (collection of ships—NT3oa is Atlantic Ocean</td>
<td></td>
</tr>
<tr>
<td>navy PI5n: ∑EO925111^PI5n &gt; NT3oa</td>
<td></td>
</tr>
<tr>
<td>Missile defense</td>
<td>Confucianism impact on modernization</td>
</tr>
<tr>
<td>355.45(075.8)</td>
<td>22</td>
</tr>
<tr>
<td>3 Social science—35 Public administration—355</td>
<td>2 Religion—21/29 Religious systems—22 Religions originating in the Far East</td>
</tr>
<tr>
<td>Military affairs—355.45 Defence of the realm—</td>
<td>CR1d→EO†</td>
</tr>
<tr>
<td>common auxiliary of form: Texts for university,</td>
<td>(Confucianism)(influences)(economic output)(growth)</td>
</tr>
<tr>
<td>higher education PI5n→ga E094612 (military)(prevents)(missiles)</td>
<td></td>
</tr>
</tbody>
</table>
We have used bold type as a simple mechanism to illustrate the presence or proximity of the concept in each case. In most cases the concept itself is not explicit in the UDC string, which is used to create discipline-based gatherings. In all cases the concept is explicit in the BCC string, and in all but one case it is the first element in the string. Note especially the clarity of the sentence-like BCC strings. The phenomenon-based gatherings, then, are more precisely conceptual.

4. Network analysis

Network analysis allows us to discover and visualize nodes representing main classes, auxiliaries, and connectors in UDC classified strings, and nodes representing classes and relations in BCC. This technique of analysis, which is explained fully in Smiraglia et al. (2013), involves construction of matrices using the quantity of co-occurring connectors; the matrices can then be used to create both multi-dimensionally-scaled plots to demonstrate co-occurrence and network diagrams to show navigable pathways and their relative strengths.

Among the UDC classified strings, 209 strings (or 54%) represented single term expressions using no auxiliaries or connectors. The majority of these strings occurred in class 3 “social sciences” (35%) and class 6 “applied sciences” (32%). Class 5, 7, 1 and 2 occupied another 27%. Analysis of two-digit disciplinary classes shows the majority of strings occurred in 37 “education,” 62 “engineering,” 33 “economics” and 61 “medical sciences.” Of course, underlying these strings is the hierarchical network of disciplines divided by class, division and subdivision, creating a tree-like structure. The navigability of a classification’s syntetic structure is well-understood, and thus we have not looked further into that in the present study. On the other hand, 144 of the UDC classified strings (37%) were combinatorial strings using main classes, auxiliaries and connectors. These strings fell mostly in classes 8 “language, linguistics, literature” (25%), 6 “applied sciences, medicine, technology” (24%), 3 “social sciences” (22%) and 7 “the arts, recreation, entertainment, sport” (11%). Analysis by two-character disciplinary codes showed that the largest clusters were in 38 “education” (23%), 82 “literature” (18%), 62 “engineering” (13%), 83 “German literature cancelled number” (13%) and 61 “medical sciences” (11%), with smaller clusters in chemistry, sport and philology. Thus, there is a small but important difference between the two groups of classified strings such that those using UDC’s faceted synthetic structure tend to be more associated with technology, education and literature. For example, in the UDC string “808.2(075.3),” main class “8 Language. Linguistics. Literature” is linked to “(0) Special auxiliaries of form.” A network map is a way of visualizing these linkages as they occur in a set of classified strings, such as the sample in our study. For these matrices were constructed and used to create a visualization of the networks underlying the combinatorial structure, shown in Figure 3.
The size of the nodes indicates the strength of the linkage. For example, classes “8” “6” “3” and “9” most often link to special auxiliaries of form or language, classes “3” “0” and “7” have most of the chronological form linkages, classes “8” and to a lesser extent “5” “6” and “7” link to simple relations. Classes “1” “7” “5” and “9” have most of the consecutive extensions, and classes “6” and “8” are most linked through addition to each other. Classes “2” and “1” have few linkages and thus appear at some distance from the core of the network.

Among the BCC strings, sixty represented single term expressions with no relators or synthesis. These fell mostly in classes T “technology and science,” P “politics,” C “culture,” E “economy,” and A “art.” There were 301 combinatorial strings, for which a matrix was created. Figure 6 shows the network map of the combinatorial structure among the BCC strings in this study. We can see first that there is a much denser network, and also that many nodes are represented by relators; this is a reflection of the grammar-like structure of the BCC in implementation. Mirroring what we saw in Figure 4 above, classes T, P and E are the most influential (reflecting the content of the sample). The strength of the connections is reflected in the density of the curved lines; for example, classes E “economics” and T “technology and science” is heavily connected to each other as well as to class S “social structure,” reflecting in this case the large number of resources relevant to kinds of workers in economics and the sciences.

5. Conclusions
This is the first large-scale empirical analysis of a phenomenon-based classification applied to a set of resources in parallel with the discipline-based UDC. Similar to the results reported in our preliminary study (Szostak and Smiraglia 2017, 13), this analysis demonstrates the greater economy of the BCC: “The present study demonstrates the
greater economy provided by the phenomenon-based BCC classification, which combines conceptual semantic representations in precise relator-defined syntactic strings”. The present study adds to that conclusion the demonstration of greater conceptual precision in the phenomenon-based BCC. The network analysis suggests that there is great navigational strength in both approaches.

Acknowledgments
We are grateful for the assistance of Hyoungjoo Park, Xin Cai and Vanessa Schlais in drawing the sample. We are in great debt to Aida Slavic, Editor-in-Chief of the UDC for assistance with cancelled sections of the UDC.

References
Culturally relevant relationships: publishing and connecting digital objects in collections of archives, libraries, and museums over the Web

Abstract
Cultural heritage institutions such as archives, libraries, and museums have the mission of curating and safeguarding societal memories. The patrimonialization and curatorial processes developed by such institutions are continuous value-adding activities. Such institutions are beginning to publish their collections as digital objects on the Web. The representation and publication of cultural heritage records as digital objects using LOD—linked open data—technologies is a new step in patrimonialization and curatorial practices. Many collections are thematically superimposed, complementary, and have synergy. These collections often present culturally relevant relationships between their objects, like a book about a painting or a draft or sketch of a famous painting, etc. This research aims at characterizing such culturally relevant relationships, compiling an inventory, and organizing them in a vocabulary. The implementation of semantic links using LOD technologies can achieve interoperability between digital collections and improve the usability of digital collections, thus empowering heritage institutions.

1. Introduction

This paper reports new results of ongoing research that was first communicated at the 14th International ISKO Conference in 2016. Cultural heritage institutions such as archives, libraries, and museums have the mission of curating and safeguarding societal memories, a continuous value-adding process resulting in cultural heritage objects that comprise different collections. Such institutions are now publishing their collections as digital objects over the Web. Archival, library, and museum collections are still dependent on catalog systems and technologies that do not allow full integration of their data with other resources available through the Web.

The representation and publication of cultural heritage records as digital objects using the facilities offered by LOD technologies is a whole new step in patrimonialization and curatorial practice. LOD technologies enable direct publication of digital collections and their integration into the mainstream Web. Many such collections are thematically superimposed and complementary, with the potential for as yet unexplored synergies to be developed. LOD technological resources enable such complementarity to be empowered for the benefit of heritage institutions, culture, and education.

Digital collections often present culturally relevant relationships between their objects, such as a book about a painting, a draft or sketch of a famous painting, a letter from an author commenting on a book or an artwork, or a contract to commission a sculpture or artwork, etc. This kind of synergy invites us to question what culturally
relevant relationships may exist between the digital objects of collections in archives, libraries, and museums, and how they can be discovered and identified. This, in turn, leads us to enquire how LOD technologies can be used to implement such relationships as semantic links, and how they could be useful for art, history, or culture curators to annotate digital heritage objects.

This paper seeks to discuss the questions raised above and characterize these culturally complementary relationships. The research thus involves compiling an inventory of such interrelations, organizing them in a vocabulary, and discussing how semantic links expressing them should be derived from the databases of catalog systems. The implementation of semantic links using LOD technologies can achieve interoperability between digital collections. This study also aims to improve the usability of digital collections in archives, libraries, and museums, thus empowering heritage institutions. The paper is organized as follows: after the introduction, section 2 describes the materials and methods used; section 3 describes the assumptions made concerning precise characterization of what objects we are linking and what their digital representations are; section 4 develops a framework for analysis of the relationships gathered; section 5 presents the relationships found; and section 6 provides final remarks and conclusions.

2. Material and methods

Bibliographic and document sources about the patrimonialization and curatorial processes developed by heritage institutions such as archives, libraries, and museums, were sought to supply definitions of archives, collections, items, records, and cultural heritage objects. Use cases or examples of relationships between objects suggested by curators of archives, libraries, and museums, or mentioned in literature, were collected and used as examples of possible relationships. Conceptual models such as the FRBR, the CIDOC CRM, the EDM, the RiC-CM were also examined as sources of possible relationships between objects.

A framework to analyze and organize the collected relationships was also developed, based on the top-level relationship schema between entities of Groups 1, 2, and 3 of the FRBR model. This choice is because the FRBR model is primarily object-oriented (it is concerned with relationships between objects in library collections), while the CIDOC CRM and EDM are mainly process-oriented. A deductive process based on this framework combined with an inductive process based on the cases collected were used to find the results.
3. Assumptions

It is important to define the precise nature of the objects in collections of archives, libraries and museum that we intend to link. According to Van Mensh (1992, 67), “The museum object is considered to be the basic unit of the museum working procedures”. We assume that archives and libraries wish to integrate their collections with heritage objects.

Web access to collections of heritage objects presupposes their representation in digital formats. The digital objects that are published and interlinked through the web using LOD technologies are indeed artefacts, even if the original object they are based on is a natural object (Marcondes et al. 2016). In this sense, they are social creations (Searle 1995), knowledge tools created based on archives, libraries, and museum methodologies and standards. They are complex digital objects, here called HOs – digital heritage objects – identified by a unique persistent identifier, along with metadata that provides context, access points, and enables their management in the digital environment. These metadata sets are associated with digital images or copies of the physical object. Such objects are of a priori cultural relevance, as they are the result of curation processes developed by heritage institutions.

4. A framework to analyze relationships between cultural heritage objects

A framework to analyze the suggested relationships was developed. It consists of a table cross-relating entities according to the type of heritage institution: archives, libraries and museum heritage objects. To these HOs “monuments” were also added, as there are several suggested use cases that include relationships between archives, libraries, and museum objects with monuments. Such entities are related to other entities, namely Agents (FRBR Group 2 Entities), Concepts, Processes, Time and Place (FRBR Group 3 Entities). The table is shown below. Due to space limitations, the relationships between HOs and Agents, Concepts, Processes, Time and Place, and a classification schema are not included in this paper.

Table 1: Relationships between heritage objects

<table>
<thead>
<tr>
<th>aHO</th>
<th>lHO</th>
<th>mHO</th>
<th>monHO</th>
<th>Agents</th>
<th>Concepts</th>
<th>Events</th>
<th>Time</th>
<th>Place</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>26</td>
<td>27</td>
<td>28</td>
<td>29</td>
</tr>
<tr>
<td>31</td>
<td>32</td>
<td>33</td>
<td>34</td>
<td>35</td>
<td>36</td>
<td>37</td>
<td>38</td>
<td>39</td>
</tr>
<tr>
<td>41</td>
<td>42</td>
<td>43</td>
<td>44</td>
<td>45</td>
<td>46</td>
<td>47</td>
<td>48</td>
<td>49</td>
</tr>
<tr>
<td>Agents</td>
<td>51</td>
<td>52</td>
<td>53</td>
<td>54</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concepts</td>
<td>61</td>
<td>62</td>
<td>63</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Events</td>
<td>71</td>
<td>72</td>
<td>73</td>
<td>74</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>81</td>
<td>82</td>
<td>83</td>
<td>84</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Place</td>
<td>91</td>
<td>92</td>
<td>93</td>
<td>94</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Each table cell may contain a direct relationship from the entity represented by the specific line to the entity represented by the specific column, i.e., a semantic link. The entity in the cell line is the domain, and the entity in the column is the range of the relationship. For example, cell 22 represents the FRBR Group 1 work-work, work-expression, work-manifestation, and work-item relationships. Since we asked for heritage institution curators to suggest cases of relationships between objects in collections of different heritage institutions, the framework developed reflects relationships where domain and range are objects in archives, libraries, and museum collections. Archives, libraries, and museum objects may be further characterized into different types of objects as is usual in archive, library, and museum collection management. All relationships are also types of associative relationship, largely used in thesaurus theory and construction.

“Dependence” is a fundamental criterion to analyze and classify relationships. Are there several types of dependence? Following Guarino and Welty (2000, 2009) on “existential dependence,” and IFLA (1998, 66) on “referential” and “autonomous” relationships, we question if any of the relata in the relationships found are existentially dependent on each other, or whether any of them are dependent on one another in any other sense. Searle (1995, 8) discusses “subjective judgments,” “observer-relative features” of reality, and features that are “ontologically subjective”. This then leads us to ask whether both relata are independent, or whether any relata depend on a subjective judgment from their creator or from a third-party agent, such as a curator or a literary critic. Hessen (2000) notes that knowledge is always knowledge of something; it is a relation between an agent and an object, whereby the agent is intended for the object. Within Dahlberg’s (1992, 67) concept theory, among the formal relationships, there are intersections of relationships such as those linking objects that share at least one property.

To analyze and evaluate possible relationships provided by use cases or collected in literature, each relationship is assigned a label and described, while examples are given and the following questions are asked: Do any of the relata existentially depend on each other? Is there an inverse relationship? Are there other types of relationships between the two types of objects? Are there similar relationships in other conceptual models, vocabularies, or ontologies?

5. Main results

For the purposes of this work, culturally relevant relationships are seen as those conceptual relationships that contextualize and enhance the cultural comprehension of a heritage object. They may be direct relationships, such as between a book and a painting inspired by it (e.g., the work Don Quijote de La Mancha and the aquatint by
Picasso portraying Don Quijote and Sancho Panza\(^1\), or indirect ones, such as between a book or a painting and its author or subject (the FRBR Group 1 relationships to Group 2 and 3 entities).

Such relationships may be directly derived from records in catalogs, for instance between two works with common properties, such as the title in the previous example of Don Quijote, or between a work and its author. Yet they can also be *authorial*: different cultural experts and curators, such as art and literary critics, historians, educators, journalists, scholars, etc., discover, illuminate, evaluate, relate to, interpret, and show different points of view about historical facts or processes, historical characters, and artefacts, etc. While doing their job, these experts may find or propose authorial relationships between such entities not previously perceived by anyone else.

5.1. Criteria for analyzing relationships

From the theoretical bases used and from this inventory of relationships, an initial set of criteria was developed for organization of the relationships found. Such criteria are something like what Guarino and Welty (2000) call “meta-properties”.

- **Cultural association (CA)**: when there is a relationship between two HOs established not by the creator of any of them, but by a third-party agent, for example, by a curator, a literary or art critic. Cultural association means that the two HOs are existentially independent.

- **Cultural dependence (CD)**: when two HOs have a relationship established by the creator of one of them; the creator of one HO *intended for* the other HO; the two HOs are both artefacts.

- **Cultural independence (CI)**: when two HOs have a relationship established by the creator of one of them, the creator of one HO *intended for* the other HO, but only one of the HOs is an artefact, whereas the other one is originally a natural object.

- **Existential independence (EI)**: when the two HOs exist independently of each other.

- **Intersection (IS)**: both HOs share at least one common property.

Another criterion that seems to define the way two HOs are related is the type of *expression form*, or how each object is expressed or manifested. In this sense and according to previous research (Marcondes *et al.* 2016), the following types of expression forms exist:

- (originally) natural objects
- artifacts

\(^1\) See https://www.moma.org/collection/works/68157.
– image
  ▪ iconographic (paintings, drawings, etchings, photographs)
  ▪ moving image
– textual
  – one-copy textual (documents, letters, deeds)
  – various-copy textual (books, manuscripts or print copies)
– sound
– objects (three-dimensional objects).

5.2. Relationships identified

A based_on relationship between different types of HOs presupposes an original, previous work and another work on which it is based. It encompasses all kinds of pragmatic replicas or artistic copies, re-creations, revisits, and re-readings of a work; it is concerned directly with works in the FRBR model sense. A work is based on another if the based-on work carries at least one property of the base work: -a book (IHO) which is based_on another (IHO): cell 22; -an artwork (mHO) which is the base for another mHO): cell 33; -a monument (monHO) which is the base for another (monHO): cell 44.

– Example: many literary works are based_on Shakespeare Hamlet\textsuperscript{2}, such as Hamlet for Kids (Shakespeare Can Be Fun!) by Lois Burdett; the design of the Federal Hall in New York City is based_on the design of the Parthenon in Athens\textsuperscript{3}; the different based_on versions of Da Vinci’s Mona Lisa by artists such as Dali, Botero, Andy Warhol, etc.
– Criteria: CD, IS; both HO share the same expression form.
– Do any of the objects depend on the other? Existentially, neither objects depends on the other but (we suggest), from a cultural standpoint, the based_on works would not have existed if the original work did not exist.
– Inverse relationship? Yes, the Base_for relationship.
– Are there any other types of relationships between the two types of objects? The Design_or_Procedure_for relationship.
– Are there similar relationships? crm:P15 was influenced by (influenced) and frbr:is a transformation of, has adaptation, has an imitation relationships.

Design_or_Procedure_for relationships: -between architectural plans (aHO) and a monument (monHO): cell 14; -between an artwork (mHO) and their preparatory sketches (mHO): cell 33.

\textsuperscript{2} See https://en.wikipedia.org/wiki/Literary_influence_of_Hamlet.
\textsuperscript{3} See https://en.wikipedia.org/wiki/Federal_Hall.
- Criteria: CD.
- Example: the architectural plans of MAC Niterói and the monument itself; the preparatory sketches and *Guernica* by Pablo Picasso; the preparatory sketches and the “War and Peace” panels by Brazilian artist Candido Portinari at the United Nations headquarters in New York.
- Inverse relationship? Yes, the *Design or Procedure* relationship: cells 11 and 33.
- Similar relationships? The *crm: E29 Design or procedure entity*, used with the *crm: P69 has association with* relationship.
- Other types of relationships between the two types of objects? No.

*Documents* relationship between a field notebook (aHO) and the objects it documents (mHO): cell 13.
- Example: Darwin’s Beagle’s expedition field notebook⁴ and the species collected by him.
- Criteria: CI; the range HO has the form expression *textual*.
- Inverse relationship? The species in a museum *is_Documented_by* a field notebook: cell 31.
- Are there other types of relationships between the two types of objects? No.

*Has_as_Subject* relationships: -between a painting or drawing (domain mHO) whose subject is a book (range IHO): cell 32; -between a book (IHO) whose subject is letters (aHO): cell 21; -between a book (IHO) whose subject is a book (domain IHO); - between a book (IHO) whose subject is a monument (monHO) cell 24; -between a letter (domain aHO) commenting on or describing a book and the book itself (range IHO): cell 12; -between a letter (domain aHO) commenting or describing an artwork and the artwork itself (range mHO): cell 13.
- Criteria: CD; the range HO has the form expression *textual*.
- Example: La Joconde : essai scientifique / sous la direction de Christian Lahanier⁵, like many other books, has as its subject, or describes, or analyses, Da Vinci’s *Mona Lisa*; a letter from Brazilian writer Machado de Assis to his colleague Joaquim Nabuco commenting on the idea for a future book, *Memorial de Ayres* (Jackson, 2009, 18); the book *Brunelleschi’s Dome: The Story of the Great Cathedral in Florence* (Ross 2008) has as its subject the construction of Brunelleschi’s Dome of Santa Maria del Fiori church;
- Inverse relationship? The *Is_Subject_of* relationship between a painting or drawing (mHO) whose subject is a book (domain IHO): cell 32; or between a monument (monHO) which is the subject of a book (IHO): cell 42.
- Similar relationships? The “is the” relationship *frbr:has_as_subject*, or *edm:P120*
is about (is subject of).

**Influenced** relationship: -between a work which influenced the creation of another work according to someone.
- Example: according to several literary critics, the work *Don Quijote* by Cervantes Saavedra influences many others; any of the two HO is intended for the other.
- Criteria: CA, EI.
- Inverse relationship? The *Influenced_by* relationship.
- Similar relationships? The *dbpedia: influenced* relationship.

**Inspired** relationship: -between a book (domain IHO) which inspired a painting or drawing (range mHO): cell 23; -between an artwork (domain mHO) and a book (range IHO): cell 32.
- Example: the previously mentioned book *Don Quijote de La Mancha* by Cervantes Saavedra and an aquatint by Picasso portraying Don Quijote and Sancho Panza; or the romance *Iracema* by the Brazilian writer José de Alencar and a painting with the same title by José Maria Medeiros⁶; or the Da Vinci’s *Mona Lisa* and the novel *Mona Lisa Overdrive* by William Gibson⁷, among many others.
- Criteria: CD; the two HOs have different expression forms.
- Inverse relationship? The *Inspired_by* relationship.
- Other types of relationships between the two types of objects? The *Has_as_Subject* relationship between a book (domain IHO) whose subject is a painting or drawing (mHO) or the *Is_Illustrated_by* relationship between a book (domain IHO) which is illustrated by a painting or drawing (mHO, of type iconography).
- Similar relationships? The relationship *wikim: inspired*.

**Is_Illustrated_by** relationship: -between a book (domain IHO) that is illustrated by a painting or drawing, (range mHO, of type iconography): cell 23.
- Example: Aristophane’s *Lysistrata* edited by the Limited Editions Club, which is illustrated by six signed etchings by Pablo Picasso; James Joyce’s *Ulysses* 1935 edition illustrated by Henri Matisse’s rare etchings.
- Do any of the objects depend on the other? In this case, it seems to apply the FRBR (IFLA, 1998, 66) *referential* relationship. Matisse’s etchings are components of the Ulysses edition and they would not have existed if the edition did not exist.
- Criteria: CD, domain IHO has the expression form textual, range mHO has the expression form iconographic.
- Inverse relationship? The *Illustrates* relationship.

---

– Other types of relationships? The frbr:Has_as_Subject relationship between a book (domain IHO) whose subject is a painting or drawing (mHO); the Inspired relationship between a book (domain IHO) which inspired a painting or drawing (range mHO).
– Similar relationships? crm: P65 shows visual item or crm: P46 is composed of (forms part of).
  Portrays refers to a relationship between an artwork (domain mHO) of type iconography and a monument it portrays (range monHO): cell 24.
– Example: several paintings made by French Impressionist artist Monet portraying the Rouen Cathedral\(^8\).
– Criteria: CD; range mHO has the expression form iconographic.
– Inverse relationship? The Is_Portrayed_by relationship: cell 43.
– Other types of relationships? No.
– Similar relationships? The crm: P62 depicts (is depicted by).

Apart from the relationships between objects belonging to collections of heritage institutions, there are also external relationships, that is, those between HOs and external entities such as Agents, Concepts, Processes, Time and Place. Relationships of this kind, provided in the FRBR model, such as, the relationships between a painting by Claude Monet and the concept Impressionism art movement or between a work and its author, will be analyzed in a future paper.

6. Concluding remarks

The publishing of digital collections over the web opens up new opportunities for heritage institutions. It enhances access, enables reuse, and achieves full integration of collections to the mainstream Web, thus enlarging their reach and synergies. This synergy can be exploited by implementing LOD links to establish culturally relevant relationships between the digital objects in these collections. The interlinking of resources from different institutions provides rich contexts not available through OPAC technologies (Sanderson et al., 2017). The reciprocal implementation of LOD links between heterogeneous and distributed digital collections requires cooperation, coordination and curation activities on a new level. It can also achieve interoperability, improve synergies and usability between collections, thus empowering and reshaping heritage institutions.

References


Marcia Lei Zeng, Sophy Shu-Jiun Chen

Derivative interpretation of biographical sketches (bios) supporting innovative information

Abstract
The goal of this research is to define an effective method and framework for the derivative interpretation of online biographical sketches (bios) which will lead to innovative information access. A related objective is to reuse digital archives and digital library resources which have been established all over the world during the last two decades. Specifically, this research aims to present bios using structured data based on openly available resources (including unstructured and semi-structured data, ranging from digital archives to individual web pages). When dealing with non-centralized data sources, a unified framework representing interoperable models and a platform implementing the schemas are essential. Meanwhile, machine-processable languages enable the structured bio data created using this platform to be understandable by both human users and machines. The potential of these outcomes is beyond publishing bios on the web because the semantically enriched structured data will enable innovative information access and the reuse of digital resources across the boundaries of language, geography, format, and discipline.

1. Introduction
This research was initiated from a desire to expose information about native Taiwanese artists in an effective method and reuse digital archive resources. During the last two decades, the World Wide Web welcomed all kinds of new digital collections, domain-specific information resource portals, online exhibitions, and other products which expose and provide access to the not-born-digital resources hosted in libraries, archives, and museums (LAMs). The completion of these digital collections and services involved very complicated digitization and documentation processes with tremendous investments from governments and institutions all over the world. Now, as the world enters the Semantic Web, the advanced utilization of these digitized resources must go beyond conventional searching and browsing within the provided digital silos which are restricted by the individualized models and schemas initially applied. Aimed at conveying new levels of usefulness for these digital resources to the international community across the boundaries of language, geography, format, and discipline, this project chose to use semantically enriched structured data to share, connect, and reuse these digital resources.

Four unique points reflect methods that may distinguish the current approach from others. 1) Instead of binding tasks to a centralized project, a workflow and a platform have been developed for the derivative interpretation of available online biographical sketches (bios) that can be generated by any individual contributor. 2) Semantically enriched structured data following widely used standard schemas are used as the key to share, connect, and reuse digital resources. The produced data are machine-
understandable and search-engine friendly. 3) The bio sketch sources used in generating the structured data can be from any digital collection or individual website, from structured, semi-structured, and non-structured data that were created by any trustable source. 4) The resulting structured data can have multiple-output and reuse potentials: embedded in an existing webpage, stored as RDF datasets with a SPARQL endpoint, integrated with other products, and visualization.

2. Rationale

Using the materials from the Taiwan Digital Archives, the Academia Sinica Center for Digital Cultures (ASCDC) released a comprehensive book-style digital biography website (in Chinese and English) in 2014, *Starting Out from 23.5°N: Chen Cheng-po* (http://chenchengpo.asdc.sinica.edu.tw/main_en), for one of the most well-known Taiwanese artists, Chen Cheng-po (陳澄波, 1895-1947, also known as Tan Ting-pho). The website, featuring more than 500 pieces of artwork, correspondence, and other documents, is organized into 10 themes based on Chen’s personal history, traveling paths, and social networks. Many items in the Chen collection, as well as his painting style, reveal a strong influence of both Japanese and Post-Impressionist art, with cross-cultural encounters of Chinese culture and intensive social networks. The artist was the first to be given a dedicated website and a Linked Open Data project created by ASCDC. An International Conference on Chen Cheng-po and Modern East Asian Art History was held in 2015, indicating the importance of this Taiwanese artist.

Along with Chen Cheng-po, 14 other well-known artists are presented in the “Friends chapter” of this website (ASCDC 2014). A short bio of each artist, a list of related collections, art works and awards, and a bibliography were prepared by the researchers of Academia Sinica’s Institute of Taiwan History and the ASCDC. The website’s high quality and unique design has made it well-known and generated great interest and recognition in Taiwan. To measure the international impact, two years after the website’s release, the authors of this paper collected data in January 2016. It was found that, however, only less than half of these 14 artists had knowledge graphs accompanying the English search results provided by search engines (Google, Yahoo, and Bing). This finding led the authors to further identify possible reasons of their low-level exposure to the English-speaking world, based on three questions:

Q1. Are there other relevant digital collections and are they connected? In fact, there are very rich and well-organized digital archives which document a tremendous amount of digitized resources (including those for Taiwanese artists) through a 10-year project (2002-2012), involving hundreds of institutional contributions. The integrated resources (including the metadata, representative images, digitized original documents, and multimedia files) are accessible through an online union catalog (with over five million items) of the unified Taiwan Digital Archives (http://digitalarchives.tw/).
Additionally, enormous efforts of art foundations, artists’ families, fans, art historians, and educational units have resulted in many outstanding agent-based digital archives and digital exhibitions. There are also significant materials hosted in other digital ontologies are the following:

- **Friend of a Friend** (FOAF) is an ontology focusing on Person (a subclass of Agent), describing persons, their activities, and their relations to other people and objects (Brickley and Miller 2000-).
- **DBpedia** ("DB" for "database") extracts structured data from Wikipedia. In its **DBpedia Ontology** (2008-), the Person class has over 260 sub- and sub-sub-classes (e.g., Artist, Writer, Athlete, Skier). The Artist class has nearly 20 sub-types and about 250 properties, including many inherited from the Person class.
- **BIO** is an event-centered ontology for describing a person's life as a series of interconnected key events (assumed to occur over a period of time) around which other information can be woven. It provides types of Events and properties of a person and an event, plus properties for relating an event to an agent and to other events (Davis and Galbraith 2003-).
- The **Biography Light** model takes an event-centric approach and has reused properties from other ontologies such as FOAF, BIO, OWL-Time, GeoNames, and Basic Geo. It is aimed at prosopography (also referred as “collective biography”) in order to connect biographic events to other resources available through the Semantic Web (Ramos 2009; University of California, Berkeley 2009).
- **Schema.org** (2011-) was created by major search engines, aiming to provide many schemas under one namespace so that webmasters can describe and expose websites of any kind to search engines. Its impact can be seen from an increasing number of website products across almost all domains (the Person class claims to be used in 1,000,000 domains as of today). Schema.org managed to use a limited number of properties (>50) for Person, with inherited properties (12) from the upper-class Thing. Additionally, 60+ properties for the relationships between instances of Person with other classes are listed in the Person schema.

### 2.1. Examining the components of online bios

In order to examine the major components of online biographical sketches, the authors collected data from the following web resources: 1) two general encyclopedias; 2) 15 international websites dedicated to collective biographies of persons from certain domains, regions, or time periods; and 3) a dozen websites, each being designed for a specific individual person, ranging from historical figures to current professionals, produced by institutions and third parties or by the individuals themselves. Attention was given to four areas: the common components of online bios (e.g., narrative, timeline, fact list, picture, bibliography); structured data provided on the bio pages (e.g., full name, birth date, birth place); variations (e.g., the types of relations, roles, awards,
intellectual or artistic works); and elements usually not found in ontologies (e.g., quotes, attribution, parent occupation, etc.).

2.2 Creating and implementing an application profile

After a comprehensive review and comparative study, the authors created an application profile based on Schema.org. This decision was based on the fact that the Schema.org ontology and its recommended Microdata and RDFa encoding formats provide an interoperable model for anyone who would like to embed structured data into an online bio page for exposure to machines and search engines. An application profile is needed to handle the challenges that any creator would encounter, for instance, the appropriate application of these 100+ properties about or related to the Person class, along with the correct encoding for embedding Microdata- or RDFa-encoded data into a web page.

The application profile accommodates the necessary components to highlight information about a specific person with an interoperable model for anyone to use. A unified framework resulted in two templates for creating structured data based on the application profile. A full description’s template includes properties in three sections: 1) Basic Information about the person, 2) Relations (family, sociality, and others), and 3) Important Roles. This last section uses various properties that are affiliated with other ontological classes (such as CreativeWork, Organization). Examples of roles include: as an author, composer, director, producer, or contributor of a creative work; as an actor or performer of a creative work or event; as an organizer, founder, or owner of an institution; as an award receiver; and as a character in a creative work. The full version fulfills the need of describing people who played important roles in history. The short version covers the basic information about a person that are typically useful for online resumes and short bio sketches of professionals.

2.3 Developing a workable platform

From the perspective of the contents of a biography, the semantic meaning of a particular fact in an online bio should be encoded so that search engines and other computer programs can understand, react, and process these structured data that are embedded in a bio. Two basic encoding formats, RDFa and Microdata, function as a means of nesting metadata within existing content on webpages to allow search engines, web crawlers, and browsers to extract and process the structured data from a webpage. To achieve the result of interactive encoding and generate desired source codes on-the-fly, the project designed interactive web forms as the communication channel between a human editor and a computer. As one enters the information about a person (left in Figure 1), such as the person’s name, related URL, nationality, birth date, important roles related with creative works, and relationships with other persons and organizations, the structured data are generated simultaneously (right in Figure 1),
with the coding either in Microdata or RDFa syntax.

The template not only performs the basic fill-in functions, but also automatically embeds a resource’s Uniform Resource Identifier (URI) with the identity of a “Thing,” e.g., a person, a creative work (painting, book, video), or an institution (university, society, association). In cases where the property is designed for a class other than “Person,” inverse relationships would be generated. The form is designed to be able to expand when dealing with multiple relationships between a person and other things.

Figure 1: An interactive encoding form generating desired source codes on-the-fly

2.4 Testing the workflow and the tool

The tests were conducted at three stages: 1) At the pilot project, a set of 14 Taiwanese artists’ bios available on the Chen Cheng-po website were used as the base for generating structured data, testing the templates, and establishing useful structured datasets. The data values were taken from the bios in two ways before they were encoded: human analysis of the text and machine-aided entity extraction. The results indicated pros and cons for each and proved the benefits of their combined use. 2) To extend to other domains in the humanities, bios of a set of well-known Taiwanese creative agents (writers, choreographers, and musicians) were included. Sources for the bios were broadened to any available websites related to the agents, special collections, and event announcements of performances and exhibitions from other countries. The platform’s full version template was enhanced after this stage. 3) Two dozen invited users in America were invited to perform self-tagging using the short version template. URIs of personal social media accounts were proved to function well for people who have no website or name authority entries. Visualization of the triples using RDFa-play
tool was proved to be valuable. User-desired additional properties were collected in order to enhance the short version of the application profile.

3. Preliminary results

The results for the 14 artists’ structured data were presented with webpages first. The webpages' source codes are machine-processable, embedded with either Microdata or RDFa based on one’s needs, using the Schema.org vocabulary. The source codes of a bio page constructed using this platform can be taken to the RDFa-Play tool to obtain a visualization and raw data result. For the Microdata-format, one can use the Google Structured Data Testing Tool to check how the encoded structured data in the source code (left in Figure 2) are read by a search engine (right in Figure 2). The results are very encouraging, for example, even awards were recognized.

![Figure 2: Results from Google Structured Data Testing Tool](https://search.google.com/structured-data/testing-tool)

The English bios and structured data values were submitted to Wikipedia and Wikidata manually using a mapping process. A re-examination after two to four weeks of publishing the English bios and structured data found that all of the 14 artists now have knowledge graphs on Google with varying degree of detail. The “People also search for” results have been found to reveal other artists.

A by-product of the project is a cluster of datasets of the structured bios. The resulting raw data formed by the RDFa-Play tool uses one of the RDF serializations
called Turtle. The automatically generated RDF dataset for each bio can be stored in a graph database using an open source tool such as FUSEKI, a SPARQL server. Including as many bio datasets as one likes, loads of complicated questions can be answered by querying the datasets with SPARQL query language. The datasets cluster forms a knowledge base and can be enriched greatly when linked with other knowledge bases and reused by anyone in the world.

Challenges and unanswered questions have been identified when dealing with historical figures, non-English text, and the inverse relationship’s coding. For encoding bios using Schema.org, there are obvious limitations. Because these artists are mainly historical figures, English present verbs that are used as property names (such as “follows” and “owns”) and those that deal with present situations (such as “jobTitle” and “workLocation”) look odd, even after the obvious irrelevant ones (such as “email” and “taxID”) were excluded. The contents themselves presented other problems, such as the “nationality” of persons who were born in the Qing Dynasty or during the period that Taiwan was under Japanese rule. Other problems include the handling of inverse relationships, the non-English text in persons’ original names, the punctuations in text strings, and so on.

4. Conclusion

The approach of using un-centralized resources to generate datasets has special meaning to digital archives and digital libraries in this current stage, because digital resources can be further used and enriched without waiting for a large project, or one-time effort, which would be difficult to sustain after funding ends. The derivative interpretation of biographical sketches could effectively support innovative information access. The accumulated efforts will have no limits for those who contribute, while the outcomes can be maximized. The authors understand that wider participation will happen only if the tools (such as the online bio platforms and datasets builders) are handy to use and the roadmaps are clear to follow; therefore, the interoperable framework enables great potential for wider participation and for innovative bio information exploration. Further than making it easy to tag a bio and to create machine-understandable structured data, the project demonstrates a path toward building knowledge bases. In addition to search engines, any digital library, portal, or website that is interested in contents related to and about individuals can use semantically-enriched structured data that have not been well-exposed or connected before.
References


University of California, Berkeley, Electronic Cultural Atlas Initiative (2009). Bringing Lives to Light: Biography in Context. Available at:
Louise F. Spiteri

Extending the scope of library discovery systems via hashtags

Abstract
Three cases studies were conducted to explore the potential contributions of hashtags to library discovery systems. Searches for #canpoli, #asist2016, and #solareclipse2017, and their equivalent Library of Congress Subject Headings were examined in social media sites and three library discovery systems. The results of the case studies suggest that hashtags can link library discovery systems to information resources in a variety of social media services. In an expanding linked data environment, and considering their growing popularity and scope, hashtags in library discovery systems are a new frontier worthy of exploring.

Introduction
Social tags can add value to library discovery systems by (a) enhancing controlled vocabularies; (b) providing de facto subject access to items in the catalogue to which few, or no, subject headings have been assigned; (c) adding user expertise to the description of content; and (d) enhancing readers' advisory services (McFadden and Venker Weidenbenner 2010; Mendes, Quiñónez-Skinner and Skaggs 2009; Spiteri and Pecoskie 2016; Spiteri and Tarulli 2012; Tarulli and Spiteri 2012). One of the possible limitations of library discovery systems is that, for the most part, they provide access to only items that are owned by the library system. In a world of linked data and the Semantic Web, however, does this inward focus continue to work well? Can we provide an additional service to clients by linking them to related content that is not owned by the library system? Through an examination of relevant literature, as well as three specific case studies, this chapter will explore the potential contributions of hashtags to library discovery systems.

Literature review
Chang and Iyer (2012) suggested that linking library resources to popular or topical hashtags can add value to information discovery. Users interested in learning more about a topic via a hashtag could find library materials of which they were previously unaware, rather than be directed to only Internet-based resources. In this way, hashtags serve to not only expand the scope and depth of information available to searchers, but also to promote library resources and services. Hashtags are growing in popularity, and are ubiquitous in our society, appearing in many venues, advertisements, sports events, and so forth. Hashtags allow users to self-catalogue content, and also to filter which and whose content to see.

User-generated content such as tags and, by extension, hashtags, reflects the contexts, language, and interpretations of the users (Spiteri 2012). Hashtags could thus
be a form of needs-oriented indexing, with the added advantage of not requiring library professionals to anticipate user needs. User tags can contribute to library discovery systems by reflecting the perspectives and language of the users, and by serving as access points to the content of the work, including its affect, for example, readability, tone, mood, and so forth (Pecoskie, Tarulli and Spiteri 2014).

Sharma (2013), and Romero, Meeder, and Kleinberg (2011) discussed the concepts of the *stickiness* and *persistence* of hashtags. Stickiness refers to the level of diffusion of a hashtag, that is, how many people will use the hashtag; the higher the stickiness, the greater the diffusion of information. Persistence refers to for how long and widely a hashtag will continue to be used. Hashtags that are sticky and persistent could result in greater long-term exposure to library resources.

A number of studies have examined the use of Twitter in scholarly communication. Shiffman (2012) argued that scientific conferences bring together countless experts on any given topic, but are not an effective means of reaching the interested public on a large scale. Ferguson *et al.* (2014) found that Twitter, and associated conference hashtags, assisted in sharing ideas and disseminating the findings and conclusions from presenters at the conference, and concluded that researchers and clinicians should consider using this technology to enhance timely communication of findings. Numerous scientists actively use Twitter for public engagement and outreach (Winkless 2013; You 2014). Hashtags associated with conferences can help promote discussion and debate, and new research (Darling *et al.* 2013). Providing official conference hashtags in library discovery systems can serve as an important way to link various conference related activities in one place.

As library discovery systems continue to add variety to the information resources to which they provide access, particularly in a linked data environment, where an institution's collections can range widely in scope and breadth, it may become increasingly important to tap into the knowledge and expertise of our users, particularly in the wake of shrinking budgets and the reduction in the amount of in-house original cataloguing that takes place.

Bruns and Moe (2014) presented a threefold conceptual model that examines the micro, meso, and macro layers of communication in Twitter. Micro levels comprise tweets in individual accounts. Meso levels comprise the follower/followee networks. Macro levels are hashtagged exchanges. Hashtags could create macro levels of communication via the library discovery system, connecting users both within and without a particular library community to a shared topic.

**Methodology**

Three case studies were conducted to examine the possible contributions of hashtags to library discovery systems. Three hashtags were chosen that met the criteria of
popularity, stickiness, and persistence:

- **#canpoli** is broadly used to discuss matters pertaining to Canadian federal politics. Further, there is consistent use of related hashtags for the provincial and territorial levels of government; for example, #onpoli refers to the province Ontario, #ykpoli to the Yukon Territory, and so forth. The #canpoli hashtag could provide some valuable resources related to political developments, events, and public engagement in Canada that might not be available via a library discovery system.

- **#asist2016**, refers to the 2016 conference of the Association for Information Science and Technology (ASIST), a scholarly association that meets yearly. The #asist2016 hashtag could provide some valuable resources related to activities, events, and topics associated with the 2016 conference that might not be available via a library discovery system.

- **#solareclipse2017** refers to a very specific event that garnered a great deal of coverage around the world. This hashtag could provide some valuable resources related to this solar eclipse, especially in the form of photographs and videos of the event, that might not be available via a library discovery system.

For each hashtag, an initial search was conducted in Google to examine the scope of coverage in various social media services that use hashtags, such as Twitter, Facebook, Pinterest, YouTube, and Instagram. For each concept represented by the hashtags, a search was conducted in two public, and one academic, library discovery systems: Edmonton Public Library (#canpoli); WorldCat (#asist2016); and Ottawa Public Library (#solareclipse2017). These systems were chosen because they allow for the use of social tagging, and because of their popularity amongst public and academic libraries across Canada. Because neither of these discovery systems uses hashtags, the concepts expressed by the hashtags were searched via the most-closely corresponding Library of Congress (LC) Subject Heading, as follows:

- **#canpoli**: Canada-politics and government
- **#asist2016**: Association for Information Science and Technology
- **#solareclipse2017**: Solar eclipses

**Findings**

**#canpoli/ Canada-Politics and government**

The Twitter search results for #canpoli include the opinions of members of the public about members of Parliament, the Prime Minister, or specific federal government actions. The Instagram results for #canpoli provide links to images featuring political events, rallies, politicians, and so forth. Most of the Facebook posts that use the #canpoli hashtag provide links to relevant newspaper articles and videos pertaining to Canadian politics. The YouTube results were not helpful, as few of the
videos retrieved from the search were related to Canadian politics. Very few of the videos in the results contain the #canpoli hashtag, or any hashtag, for that matter, even though this feature is available in YouTube. A very interesting result was the site Politwitter (http://politwitter.ca), which provides a site page that links you to sources about Canadian politics in different sites: Twitter, Facebook, Google+, News, Blogs, and Hansard. Of interest is the inclusion of tweets and resources that pertain to provincial developments, as hashtags such as #onpoli, #bcpoli, and #nspoli are included in some tweets. This shows that one hashtag can open the breadth of the results to include potentially relevant discussions pertaining to other levels of Canadian government. The link to this site would be a valuable addition to any library discovery system, as it provides a portal for all Twitter-related discussions related to federal, provincial, and territorial governments.

The search for the LC heading in the subject field of the Edmonton Public Library retrieved 350 items that pertained to works about Canadian politics, satire, memoirs, and biographies. The addition of the #canpoli hashtag, and particularly if it is linked to the Politwitter site, would be a valuable addition to any library discovery system, as it provides a portal for all Twitter-related discussions related to federal and provincial governments. Similarly, the #_poli hashtag pertaining to Canadian provinces and territories could be added to library resources that focus on these topics.

#asist2016/ Association for Information Science and Technology

The Twitter results for #asist2016 are varied in scope. Many tweets discuss topics mentioned in presentations. There are pictures of events, links to outside papers that support the topics discussed, blog posts discussing the conference from participants, and so forth. Many of the tweets include direct links to named researchers (via @), which allows people to associate events and topics with specific researchers, as well as to follow each other. The tweets provide a historical archive of the conference that exposes people who could not attend the conference, or certain sessions, to many of the discussions and themes that occurred. More importantly, perhaps, are the @ direct links to the researchers associated with the conference presentations, as this opens the door to expanded networks of communications. These tweets and links provide useful supplements to the official published conference proceedings. The Google search retrieved links to individual blog posts written by academics who discussed their experiences at the conference. The Facebook results for #asist2016 included many official notes and links from the ASIST Facebook page, as well as commentaries, photographs, and links to relevant articles from attendees of the conference. The Instagram page retrieved 40 posts of photographs posted by attendees at the conference. There were no relevant results in Pinterest or YouTube.

A search for the LC heading in the subject field of WorldCat yielded 91 results that were not particularly useful, as they pertained mostly to published directories of various
associations, among which ASIST was listed. The use of the author field resulted in 56 items which pertained to various ASIST publications, including the *Journal for the Association of Information Science & Technology*, and conference proceedings. The conference hashtags would be a very useful addition to WorldCat, as they provide a snapshot of the various conference activities, links to scholars who presented, the musings of academics about the presentations they attended, and so forth. In this case, it would be useful to link only to the Twitter page for each year's conference, as it provides the richest source of material.

**#solareclipse2017/Solar eclipses**

The Twitter results for this hashtag provide links to many photographs of the eclipse by both professional and amateur photographers in different parts of the world. There are some other tweets that pertain to people's reactions and discussions, but the most notable content of this Twitter feed are the photographs. The Instagram results provide a large variety of images of the eclipse, mostly by amateur photographers, as well as videos; this is true also for Pinterest and YouTube, where many posts included the #solareclipse2017 hashtag, which is in marked contrast to the results for the other two hashtags. The Facebook results provide many photographs, memes, and personal reactions, as well as links to live streams of the eclipse from organizations such as NASA.

The search for the LC heading in the subject field of the Ottawa Public Library provided 11 results: Seven of the items pertain to non-fiction works about the phenomena of social eclipses, while three items pertain to works of fiction that involve solar eclipses. The #socialeclipse2017 hashtag is particularly interesting because it provides a snapshot of a very specific event in time. This hashtag could make a valuable addition to library discovery systems, and could serve as an important archival tool to document events, photographs, and videos of the 2017 solar eclipse, and which may not be available via the library catalogue.

**Discussion & conclusion**

Hashtags have the potential to make positive contributions to library discovery systems; it is necessary, of course, to consider possible detractions and ways in which to mitigate them. Library discovery systems, such as WorldCat and BiblioCommons, have allowed users to add their own tags to metadata records for a number of years, so why consider hashtags? Social tags focus only on items owned by the library system; as such, they function in a similar manner as Library of Congress Subject Headings. While they serve also to describe aspects of information resources, hashtags differ from social tags in their ability to link to outside sources of information, such as Twitter, Facebook, Instagram, Pinterest, and so forth. As we saw in the searches conducted for the hashtags #canpoli, #solareclipse2017, and #asist2016, hashtags can extend the
resources available to users via the library discovery system.

The trending nature of hashtags can be both a blessing and a curse. On the one hand, hashtags may be better capable of reflecting new and emerging topics than, say, standard library indexing systems, such as the Library of Congress Subject Headings (LCSH). So, for example, the concept expressed by the #MAGA (Make America Great Again) hashtag, cannot yet be expressed well by LC headings, as the editorial approval processes for the creation of new subject headings can be lengthy. Further, trends may be here today and gone tomorrow, so the value of some hashtags may be very short lived. With regard to library discovery systems, the stickiness and persistence of hashtags could be important factors to consider and, possibly, to measure. Hashtags that are sticky and persistent, such as the hashtags #canpoli and #asist2016, could result in greater long-term exposure to library resources. Hashtags can serve as an important archival record of the various conversations, events, and information sources related to specific events, as we saw with the #solareclipse2017 search.

Conference hashtags could be very useful in an academic library catalogue. The addition of the conference hashtag to the library discovery systems can link users to varied discussion threads and conversations pertaining to that conference; further, they can connect with various scholars, follow their work, and so forth. Conference hashtags can serve also to connect people to conferences that occur yearly, since there is often an attempt to have a common hashtag template for a conference. The library discovery system could provide all these hashtags, for example, in the bibliographic record for the journal of a society (e.g., Journal of the Association for Information Science and Technology), which would provide users with a central place to find conference information related to the society.

One can certainly understand some reluctance on the part of library staff to include hashtags in the library discovery system, as hashtags can be very uncontrolled, and could also lead people to problematic external resources. What we could have, however, is the selective addition of hashtags to test the waters, such as official hashtags for conferences or events. Thus, the selection of hashtags could be restricted to only library staff to minimize the noise that could be caused by irrelevant or inaccurate hashtags.

Hashtags can provide other voices in the library discovery system. Subject headings assigned by library staff represent one perspective and outlook that may not match that of the library users. Hashtags can thus allow for divergent views and interpretations. Serendipitous discovery, as well as the creation of environments where people can form connections with like-minded people or, perhaps even more exciting, form connections with people with different perspectives, can have their own benefits. Hashtags do not create strict ties but, rather, open opportunities for new avenues of linked concepts. These avenues might not last long, but they offer people the chance to explore and
follow them. So, rather than rely on library staff to make those connections, which can be a flawed process, of course, we give our users the opportunity to create their own classificatory networks around, say, a particular information need.

Hashtags can link library discovery systems to information resources in a variety of social media services, such as Twitter, Facebook, Instagram, Pinterest, and YouTube. Through the hashtags in the library metadata record, users can expand the scope of their searches to include items that are related to library resources, such as videos, images, streamed events, and so forth. More rigorous and detailed studies need to be conducted to explore more clearly how hashtags would function in a library discovery system, and whether the benefits do, in fact, outweigh the detractions. Many library discovery systems have incorporated many user-driven social features, such as ratings, reviews, and tags. In an expanding linked data environment, and considering their growing popularity and scope, hashtags in library discovery systems are a new frontier worthy of exploring.

References


An FRBR-based approach for transforming MARC records into linked data

Abstract
In this study, a practical workflow is outlined for transforming MARC records into Linked Data according to classes and properties of FRBR and FRAD defined in the RDA Registry. Four bibliographic records of ‘Pride and Prejudice’ were selected as subject. One was in English and the others were Chinese records to illustrate the hidden relationships embedded in MARC records, including translation, version and reproduction, relationships between classes in FRBR group 1, and relationships between FRBR group 1 and FRAD for Linked Data. In total, the proposed workflow is composed of three stages and five tasks as follows: data preparation and selection (comprising two tasks, i.e., changing MARC tags into semantic labels with their data and selecting significant labels and their data for Linked Data), data modeling (comprising two tasks, i.e., aligning MARC labels and their data with semantic equivalent classes defined in the RDA Registry and building the relationships between selected classes through the properties of the RDA Registry), and reusing existing value vocabularies for external linking (comprising one task, such as linking terms from VIAF, TGN and so on). Issues related to de-duplication and collaboration of LD are also addressed.

1. Introduction
Traditionally libraries have provided access to scholarly information resources for scientific research and communication. Most of the accessible scholarly resources owned by libraries are represented by complex structures and attributes in a Machine-Readable Cataloguing (MARC) format, MARC having been used as a means to interchange bibliographic and authority records across various information systems within the library community. Although a huge volume of MARC records are valuable in the bibliographic universe, most are not included as part of the Google search engine which has limited their direct use by netizens on the Internet. Recently, Linked Data (LD) has become the preferred approach for the conversion of MARC-based legacy records into a part of the semantic web by libraries. Based on the principles of LD, MARC-based catalog records can be sliced into LD and then be aggregated with other external resources and their contexts. With the advancement of LD, official documents released by W3C (Hyland, Atemezing and Villazón-Terrazas 2017; Hyland and Villazón-Terrazas 2011) can be regarded as useful best practices for authoring and publishing LD.

With the application of the semantic web, the Functional Requirements for Bibliographic Records (FRBR) has been regarded as a conceptual reference model (CRM) for the bibliographic universe. Many projects and studies have argued that FRBR can feasibly be used as a CRM for the semantic web. On the other hand, the entities and relationships defined by FRBR and Functional Requirements for Authority
Data (FRAD) have been implemented as classes and properties into the RDA Registry (RDAR, http://www.rdaregistry.info/) with permanent namespace and identifiers (RDA Steering Committee, 2016); however, there is still a lack of common agreement on a best practice for the implementation and use of FRBR and FRAD in the library community. Therefore, a question worth investigating is how classes and relationships of FRBR and FRAD defined in the RDAR can be employed as a CRM to transform existing MARC records into LD in terms of workflow for LD generation.

2. Literature review

LD has become an accepted approach for combining related information with diverse viewpoints for the same resource. Libraries have taken advantage of LD to free MARC-based catalogue data so that they can enrich their contextual content with cultural heritage from the Internet. Although best practices have been released by W3C (Hyland and Villazón-Terrazas 2011; Hyland, Atemezing and Villazón-Terrazas 2017), it is still a challenge to know how to move from theoretical discussion into practical implementation of LD for libraries (Hanson 2014). Libraries need to know how LD are actually created and published in practice (Hanson 2014), and be able to understand many unexpected issues (Bowen, 2010), such as linking to the exact URI (Cole et al. 2013) and workflow (Bowen 2010; Cole et al. 2013; Di Noia et al. 2016; Hallo, Luján-Mora and Trujillo 2014; Hanson 2014; Lampert and Southwick 2013; Southwick 2015).

In addition to BibFrame, FRBR is another popular CRM for LD adopted by libraries. FRBR, FRAD and Functional Requirements for Subject Authority Data (FRSAD) have been used for various applications, including as a design basis for the RDA and a CRM for data modeling during conversion of MARC into LD. In the cases of the National Polytechnic School in Ecuador (Hallo, Luján-Mora and Trujillo 2014) and the LIBRIS in Sweden (Malmsten 2008), FRBR group 1-2 has been employed, whereas in the case of the Bibliothèque Nationale de France (BNF) (Simon et al. 2013; Wenz 2013) and the National Library and Archives of Iran (NLAI) (Eslami and Vaghefzadeh 2013) FRBR 1-3 group has been used. Furthermore, the Biblioteca Nacional de España (BNE) has adopted FRBR, FRAD and FRSAD to delineate the relationship between LD embedded in bibliographic and authority data (Vila-Suero, Villazón-Terrazas and Gómez-Pérez 2012). During transformation, either data mapping (e.g., BNE) or data alignments (e.g., BNF, NLAI) were utilized to select or link to exact URI or terms from CRM classes or metadata elements. In 2014 the RDA Steering Committee (RSC) built up the RDAR composed by FRBR and FRAD with unique URIs for each entity and the relationship defined by the two aforementioned CRMs (RSC 2014). Existing cases have justified FRBR and FRAD as a feasible CRM for transforming MARC records into LD. Research is still lacking on how libraries can split MARC records into LD with a practical workflow based on RDAR.
3. Methodology

In this study, four bibliographic MARC records of ‘Pride and Prejudice’, authored by Jane Austen, were selected from the WebPAC of the National Taiwan University Library (http://tulips.ntu.edu.tw) as subject: one record was in English and the others were Chinese. Hidden relationships embedded between bibliographic MARC records including translation, version and reproduction, relationships between entities in FRBR group 1, and relationships between FRBR group 1 and FRAD were also included in this study. Three components provided by Hyland and Villazón-Terrazas (2011) were chosen as a basis to develop the workflow for transforming MARC into LD and examine the related issues (i.e., LD de-duplication and collaboration) for library-oriented LD as follows: modeling, naming with URIs, and reusing existing vocabularies.

4. Results and discussion

The workflow outlined in this study is composed of the following stages and tasks: data preparation and selection (including 1st and 2nd task), data modeling (including 3rd and 4th task), and reusing existing value vocabularies for external linking (including 5th task). The first task is to change MARC tags into semantic labels with their data. KO professionals need to change MARC tag numbers and subfield codes into semantic labels based on the MARC definition. The second task is then to select significant labels and their data for LD. Not all MARC tag numbers and subfield codes are, in fact, suitable for LD. Libraries can follow the principle “of interest” (Hyland and Villazón-Terrazas 2011) to select appropriate data from the significant MARC labels for LD. At the second stage, the most important task is to embody conceptual entities/classes and their relationships/properties in harmony with the classes and properties defined by the RDAR. The third task is, therefore, to use a crosswalk, or alignment table, to determine the semantic equivalence of selected labels regarding their MARC data and RDAR classes. Based on the results of the third task, the fourth task is to build relationships between selected classes through their RDAR properties (shown as Fig. 1). Furthermore, real instances of MARC labels and their data have been added to the aforementioned classes and their RDAR properties to validate the appropriateness of classes and their relationships. The final task is to reuse existing value vocabularies for external linking (as shown in Fig. 2). The purpose of this task is to enable library LD to become part of global LD on the Internet, and to expand the reuse or consumption by others as part of the knowledge graph generated, which is based on LD and related links.
5. Discussion

It is not an easy task to select an exact URI to distinguish individual LD from one another, simply based on URI, lexical meanings or appearances with insufficient information. Indeed, knowledge organization (KO) professionals need to perform a
series of knowledge checks with more related background information. Although LD is a useful basis for KO collaboration to seamlessly aggregate various external resources together, fundamental collaborative LD still relies on de-duplication and related issues, such as mapping and selection principles for reusing existing metadata element sets, value vocabularies and library datasets.

5. Conclusion

LD is a good approach for the practice of “thinking globally and acting locally” for KO. Although RDAR has paved the way for LD to be used by libraries with a combination of RDA and FRBR family members (e.g., FRBR and FRAD), a new approach for KO such as LD is also a paradigm shift. If the future trend of LD-based KO is to be distributed and aggregated over the Internet seamlessly, the formulation of new policies and guidelines is essential to obtain a common agreement on de-duplication, collaboration, mapping and selection for library-oriented LD.

References


M. Cristina Pattuelli

From uniform identifiers to graphs, from individuals to communities: what we talk about when we talk about linked person data

Abstract
Person identities in the Linked Open Data (LOD) environment are the result of a process of semantic representation that often requires a complex interplay of data association, reconciliation and interlinking. This paper aims to lay the foundation for a broader discussion on the role that linked data principles and practices play in shaping person identities and the communities that bind them. We have only begun to explore the potential, as well as the challenges, of shaping people’s identities through the many layers of semantics we can gather from diverse linked data sources. Using the domain of performing arts, jazz music in particular, as a scenario, I hope to offer insights on various aspects of the process of building person identities in the context of linked data applied research.

Introduction
As the vision of the semantic web gets closer to a tangible reality and the linked data cloud continues to expand in every direction, it is worth turning our attention to people as entities populating this new knowledge space. Person entities are indeed the crux of countless relationships and the catalysts for multiple contexts and Linked Open Data (LOD) technologies provide a powerful means to give visibility to this complex information environment. Individual and collective identities in the context of linked data development are the result of a process of semantic representation that often involves a combination of processes including name resolution, reconciliation and data interlinking. In discussing some of the aspects involved in building and interrelating person identities I move from the atomic level, represented by individual identifiers, toward integrated graph structures and dataset interlinking.

Identity markers
Starting at the most fundamental level, when native data is generated, linked data technologies provide an open convention for naming entities – essentially any type of thing to which a unique string of characters called Uniform Resource Identifier (URI) can be assigned. This method enables humans and machines to read and process discrete units of information, such as proper names, unambiguously. We can consider name resolution – i.e., the association of identifiers to names—the very first step toward constructing a person’s identity in the LOD context and the URI the identity marker of an individual.

As Gitelman and Jackson (2013) argue, data is never “raw” as it is the result of the cultural process of generation, curation and interpretation. Even apparently semantic-agnostic URIs carry meaning embedded in their syntax. At its most basic, the
morphology of a URI reveals the naming authority or namespace it is governed by. For example, the jazz artist Ella Fitzgerald is identified by the following URI in the Library of Congress Name Authority File (LC/NAF):

http://id.loc.gov/authorities/names/n83021406.

The information included in a namespace specification, in the form of elements and attributes, determines what semantics can be associated with a person entity and thus what statements can be made about that person. This also means that each identifier – and thus its namespace – inherently carries with it a set of assumptions and views of the world.

The general rules for inclusion in bibliographic name authorities, such as the LC/NAF or the aggregated Virtual International Authority File (VIAF) is for an individual to be either an author (or otherwise contributor), a subject or be part of a resource’s title. The authority records associated to a person entity, while have recently included additional properties beyond the one required for disambiguation (e.g., birth and death dates), are still semantically limited to a biblio-centric view of that entity.

New and fast rising representation practices underpinning the linked data paradigm, however, have started to disrupt and expand the author-centered descriptive model, opening up to new opportunities to express and reveal more articulate aspects of person entities and thus enabling richer and more complex identity construction.

Developed to support traditional practices of bibliographic control centered on proper names, library name authority services are broadly leveraged as sources of URIs, especially in LOD development for cultural heritage and LAMs. The outstanding contribution of bibliographic data to LOD development is indeed remarkable and primarily due to the wealth and quality of the data. These authorities fall short, however, when it comes to representing cultural contexts broader than bibliographic ones. These limitations became apparent when, a few years ago, we began a linked data project centered on the domain of jazz music. The project, called Linked Jazz1, began in 2012 with the primary goal of representing the relationships within the community of jazz musicians as recorded in oral histories from various jazz history archives. It has become a testbed for the development of various LOD applications and it is used throughout this paper as a source of examples and use cases.

One of the first steps in the development of Linked Jazz was to identify jazz artists from textual primary sources and assign them URIs. At that time, there were far fewer sources of authoritative URIs we could rely on for naming artists. Bibliographic name authorities could supply only a portion of the identifiers needed for the thousands of artists that emerged from the archival documents we were mining. More general-

1 https://linkedjazz.org/.
purpose linked data hubs (e.g., DBpedia\(^2\)) needed to be harvested in combination with domain-specific ones (e.g., MusicBrainz\(^3\)) to identify the many jazz artists who did not have a place in traditional name authorities. In a few instances, new identifiers had to be created from scratch.

As discussed earlier, the definition of an individual begins from assigning to a person name—a label made up of a string of characters—an identifier that represents a person entity. This process of identity management often includes the disambiguation of a name through its mapping to name authorities. Key to identity management is the use of the predicate `owl:sameAs`, the most common mechanism to express equivalence and reconcile name variants, as well as to co-reference identifiers from different namespaces. An example of co-referencing is shown in Figure 1 for the jazz musician Ella Fitzgerald.

Figure 1: Co-referencing example for person entity Ella Fitzgerald

\(^2\) http://wiki.dbpedia.org/
\(^3\) https://musicbrainz.org/
In this example, various identifiers are associated with the entity Ella Fitzgerald from general and domain specific linked data sources, creating multiple access points and making it possible to combine different layers of description about the same entity and view the artist through multiple perspectives.

With artists of the significance and popularity of Ella Fitzgerald, our running example, assigning URIs, even from multiple linked data sources, is rather straightforward. Entity resolution is instead problematic when it comes to lesser-known figures. This issue is particularly problematic within archival collections, where, for example, resources are often populated with names of local or less well-known people for whom authoritative or even public identifiers are not available. For those outliers who do not conform to the criteria of inclusion in common name authorities – e.g., because they don’t have recorded contributions – a different process of identity construction is required. While linked data publishing practices encourage the reuse of semantics whenever possible, new URIs can be minted for entities who don’t have one available.

This was the case of jazz musicians who had not recorded with a major label or who had not reached a level of public recognition, but whose existence needed to be accounted for as they were mentioned in source documents included in the Linked Jazz sample collection. We dealt with occurrences of people who didn’t have a match anywhere by minting them into the Linked Jazz namespace (e.g. http://linkedjazz.org/resource/Lynn_Grissett).

Determining whether or not to make an individual surface from obscurity and become part of the fabric of cultural linked data raises a number of theoretical, even existential, questions. There are practical considerations invested in the act of URIs creation including ensuring that evidential documentation is provided to justify the minting and that the naming agency takes responsibility to manage its local URIs for persistence and traceability. Issues surrounding minted identifiers and local vocabularies have just begun to be discussed in the LAM community, mostly focusing on methodology and technical development. Broader conversations are also needed to address the implications, for example, of giving digital life to marginal individual for digital archival practices, history research and models of historiography. Minting lesser-known musicians can indeed be a way to relate local special collections to the broader global archives of cultural memory. Moreover, it can be a powerful way to create bridges that link the margins to the center giving diverse narratives the possibility to emerge and be included.

Semantic expansion

A person means different things in different contexts and linked data principles and techniques make it possible to define identities in an expanded and multifaceted way
thanks to the mixing and matching of predicates from different conceptual models – schemas, vocabularies or ontologies – and through the co-referencing of multiple URIs.

As discussed earlier, each URI inherently carries the semantics enabled by its naming standard. As one of the key principles of linked data development (Berners Lee 2006), URIs have to be dereferenceable, in other words, they must resolve to an HTML page that can be read by machines and humans alike. Dereferenceable pages display a wealth of information about an entity in the form of linked data statements. When it comes to a person entity, we can look at their dereferenceable pages as a sort of identity card where an array of RDF predicates makes up a person’s profile while linking that entity to external data sources and to the entire linked data cloud. The semantic richness of such identity cards varies from source to source. Even a course comparison between dereferenceable pages for Ella Fitzgerald from common linked data sources convey a general sense of the artist’s representation breadth and depth. The power of description has been greatly enhanced by new community-driven linked data platforms such as DBpedia and Wikidata. The latter, a recent development of the Wikimedia Foundation, enables the direct contribution of data from volunteers with relatively low barriers. Recognized as authoritative and working together with other more institutionalized linked data services, DBpedia and Wikipedia are rapidly becoming powerful tools for diversifying and expanding the linked data ecosystem.

Recently, several new name standards and services have emerged with related databases representing new contexts, from ORCID for researchers and the global International Standard Name Identifier (ISNI) to the Social Networks and Archival Context (SNAC) for archives. They are greatly expanding the range of people that can now be identified in the linked data environment as well as the scope of the information associated with them. Because identity management is crucial to linked data development, name authority initiatives are actually proliferating to a point where issues of alignment and coordination among the various efforts are raising calls for joint efforts to promote convergence (Deupi and Eckman 2016).

The need for semantic harmonization and consistency for LOD properties describing people once again became evident in the context of Linked Jazz when we focused on gender data for a study investigating women in jazz. It is axiomatic to note that the types of descriptive attributes – RDF predicates – available, as well as the quality and consistency of their values, have a direct impact on how we can query person data –

6 https://orcid.org/.
7 http://www.isni.org/.
8 http://snaccooperative.org/.
any data indeed – and how we can ultimately use the data to support discovery and analysis. In other words, the possibility to computationally leverage a broad range of properties makes it possible to support new and unprecedented lines of inquiry.

A research resource for music historians and students, the Linked Jazz dataset includes a curated knowledge graph of over 2,000 musicians where their personal and professional relationships can be explored and queried. The dataset would be incredibly useful to investigate the historical role of women in jazz and help music historians answer research questions about influences, reputation and authority of women in jazz. This scenario, suggested by actual jazz historians, prompted us to prepare the dataset to support such a line of inquiry, including network analysis for gender ratio and distribution. A complex and problematic construct that defines person identities, gender is indeed a necessary attribute that all the person entities in our dataset needed to have as a preliminary condition to analyze our data through this specific lens.

What was deemed a rather straightforward task – *i.e.*, harvesting and ingesting a simple “demographic” attribute such as gender – turned out to be more complex and labor intensive than expected. At the time of the study (2014), gathering gender data from reliable linked data sources required significant effort and expertise as this data was sparse and inconsistent. Iterative cycles of data harvesting were needed and multiple rounds of revision and version control had to be performed to identify and correct errors. In bibliographic authorities, for example, gender values, an optional attribute, were often missing. A discussion is underway in the library community on issues surrounding gender as a descriptive attribute in the context of new cataloging practices (Billey, Drabinski and Roberto 2014).

DBpedia, another data source used in this study, also lacked a property for gender within its model. Instead, gender had to be discerned from the subjects associated with a person, which required parsing and extraction. Eventually, through a triangulation between multiple sources, we were able to successfully acquire gender information for 75% of our target list of artists (Pattuelli, Hwang and Miller 2017).

It is likely that if we had to replicate the study today we could rely on Wikidata for our main data source. Not only does Wikidata provide rich and consistent person-centered data, but when it comes to gender, it conveys a broader representation of gender types than any other linked data service. While the sources we relied on were limited to binary values (*male* and *female*) with the option of *unknown* to address uncertainty or missing information, Wikidata provides multiple options for gender variance. It also includes temporal qualifiers for making gender a time-dependent value taking into account transitions.
Semantic stratification

The full potential of linked data is achieved when data is interlinked. As discussed earlier, linked open data standards offer a suitable technical platform for semantic stratifications through the combination of predicates from multiple sources. The expressive limitations of an individual RDF vocabulary are largely offset by the endless possibilities for data interconnection and aggregation. This opens up opportunities for layered and multidimensional representations of individual and collective identities.

A powerful way to enrich and add complexity to the representation of entities is by combining diverse external datasets, such as linked data graphs, through data mashups. Data integration through interlinking from multiple sources provides the ultimate strategy to enhance resource discovery and move toward the creation of new knowledge. Case studies of dataset mashups are drawn, once again, from Linked Jazz to illustrate the usefulness of this approach.

A project using Performance History Data from Carnegie Hall (recently converted to LOD\(^9\)), was our first attempt at dataset integration. Through a series of Python scripts, artists who existed in both the Carnegie Hall and Linked Jazz datasets were identified. In other words, musicians who performed at Carnegie Hall and were recorded in the Linked jazz dataset were associated through the equivalence property `owl:sameAs`, bringing together properties from both sources for those shared person entities (Figure 2). Personal and professional descriptions were combined with information about performances resulting in an enhanced and integrated knowledge base which could be seamlessly queried. Technical details on the process are discussed in Sistrunk (2016, December 5).

Mashups also happened outside our context in the spirit of linked open data where “...the coolest use of your data will be thought of by somebody else”\(^\text{10}\). For example, the Australian linked data project JazzCats (Jazz Collection of Aggregated Triples)\(^\text{11}\) aggregated collections of RDF triples to trace performance history. JazzCats combines discography and granular performance data (e.g., solos including pitch, key, and chord) with interpersonal relationships derived from Linked Jazz to “bridge previously unconnected but complementary information about jazz music” (Bangert, Abdul-Rahman and Nurmikko-Fuller 2017).

The project JazzTube\(^\text{12}\), a joint initiative between the Hochschule für Musik Franz Liszt Weimar (University of Music Franz Liszt Weimar) and the International Audio Laboratories Erlangen, connected annotations of jazz solos with discographies, including artist data from the Linked Jazz dataset, to represent the network of interpersonal relationships between musicians who perform solo\(^\text{13}\).

These are just a few examples of how bringing different datasets together and


\(^{11}\) http://jazzcats.oerc.ox.ac.uk/.

\(^{12}\) http://mir.audiolabs.uni-erlangen.de/jazztube/about.

\(^{13}\) http://mir.audiolabs.uni-erlangen.de/jazztube/soloists/.
creating semantic bridges across diverse information spaces can open up an infinite field of connections between individuals, enabling new, unanticipated discovery. In a more diversified discovery context, people and their historical and social contexts can emerge and be more fully understood through a composite of perspectives rather than through a single narrative.

**Conclusion and future directions**

Building individual and collective identities in the context of the LOD representation framework is a complex, yet fascinating theme to explore. Pragmatic issues concerning the technical development are intertwined with the theoretical aspects of identity definition and inclusion. As linked data is based on a connectionist model, individuals – as other entities populating a networked environment – are shaped through the web of associations that link one to another. Starting from the naming process where a unique identifier makes a person begin to exist in the linked data space, we have progressed through the process of data interlinking where descriptive layers from multiple and diverse sources can be interweaved to create rich profiles and support complex queries. Aggregation, integration and mashups of sets of linked data result in new knowledge graphs where people and communities can be viewed through a wider lens and in different contexts.

The aim of this paper is to provide a point of departure for further discussion on this important area of investigation. As technical barriers to linked data development continue to be lowered and more research is conducted, new opportunities will arise to reflect on the complexities inherent in representing individual and collective identities, ultimately strengthening our shared understanding and helping establish best practices.

**References**


Identifying semantic characteristics of user interaction datasets through application of a data analysis

Abstract
In evaluating a decision, any fact analyzed needs to receive inputs from multiple data sources - structuring, integrating, storing, and processing collected data into an output that supports a better understanding of the fact from data, allowing new dimensions of analysis. The goal of this study is to identify the semantic characteristics of data attributes at the moment of collection, from dataset structures found in the data export interfaces of user interaction analysis tools, in Internet communication channels, and in web analytics data tools involved in scientific journal management, through the application of a process of data analysis and data modeling techniques. The research was delimited to exportable datasets available in interfaces from Open Journal Systems, Google Analytics and Search Console, Twitter Analytics, and Facebook Insights. An exploratory analysis approach was adopted to identify characteristics regarding how data are made available and structured in these data resources. Entity-Relationship Modeling concepts were applied to design and store data collected from services, resources, datasets, and attributes. In addition, the collected data was processed into another data structure, adopting the online analytical processing cube as a three-dimensional representation of elements, to facilitate analysis from different perspectives. This data analysis identified semantic dissonances in definitions of entity attributes, which may interfere with the process of developing relationships between attributes from different datasets, reducing the potential of interoperability.

Introduction
The use of data is part of the decision-making process in several fields, such as in education (Ikemoto & Marsh, 2007), industry (Reddy, Srinivasu, Rao and Rikkula 2010), management (Goodwin and Wright 2014), and science (Turban, Aronson and Liang 2004), among others.

In evaluating a decision, any fact analyzed needs to receive inputs from multiple data sources – structuring, integrating, storing, and processing the collected data into an output that supports a better understanding of the fact from data, allowing new dimensions or perspectives of analysis (Inmon 2005; Kimball and Ross 2011; Reddy et al. 2010; Turban et al. 2004).

For example, an evaluation of interactions between users and scientific contents in a publisher's web domain may be analyzed by service holders from the outputs generated in a process of collecting data regarding users’ interactions with their communication channels, structured into a data warehouse: a “[...] subject-oriented, integrated, time-variant, non-normalized, non-volatile collections of data that support analytical decision-making” with “[...] access to all information relevant to the organization, which may come from many different sources, both internal and external” (Turban et al. 2004, p. 236).
However, if data are analyzed as a set of elements formed by the triad of entity, attribute, and value (Santos and Sant’Ana 2015), this means using aggregated information in these elements to assure minimal semantics to understand what is available, particularly in regard to steps in obtaining data collected from data sources (Sant’Ana 2016; Turban et al. 2004).

This effort to bind information in these data elements seeks to minimize semantic dissonance between data, at the moment of data collection (Berg 2015; Rathod 2006; Ross Parry, Nick Poole and Jon Pratty 2008) – the research problem of this study.

Our aim is to identify the semantic characteristics of data attributes at the moment of collection, from dataset structures found in the data export interfaces of user interaction analysis tools, in Internet communication channels, and in web analytics data tools involved in scientific journal management, through the application of a process of data analysis and data modeling techniques.

The research was delimited to exportable dataset structures, found in journal publishing systems, online social network statistics, search engines, and web analytics tools.

The sample was restricted to dataset structures available in reports from Open Journal Systems¹, Google Analytics², Google Search Console³, Twitter Analytics⁴, and Facebook Insights⁵. These resources did not present any version control numbering on their interfaces, with the exception of Open Journal Systems (version 2.6). The data was collected in September 2017 from "Electronic Journal Digital Skills for Family Farming (RECoDAF)" accounts.

**Methodology**

An exploratory analysis approach was adopted to identify characteristics regarding how data are made available and structured in these data resources, contemplating a systematic description process for information from datasets, entities, and attributes related to interaction between users and communications channels of a scientific journal.

A total of 255 exportable datasets were found, distributed in 5 file formats: Comma-Separated Values (CSV) (82 datasets), Google Docs Spreadsheet File Format (69 datasets), Microsoft Office Open XML Format Spreadsheet file (XLSX) (50 datasets), Portable Document Format (PDF) (50 datasets), and Microsoft Excel Binary File Format (XLS) (3 datasets).

---

¹ Open Journal Systems is an open-source software developed by Public Knowledge Project, under GNU General Public License.
² Google Analytics is a web analytics service by Google LLC.
³ Google Search Console (formerly known as Google Webmaster Tools) is a web service by Google LLC.
⁴ Twitter Analytics is a web analytics service by Twitter, Inc.
⁵ Facebook Insights is a web analytics service by Facebook, Inc.
The 82 CSV datasets are distributed on 5 services, 50 retrieved from Google Analytics, 20 from Google Search, 7 from Open Journal Systems, 3 from Facebook Insights, and 2 from Twitter Analytics.

Except for CSV, all other file formats were discarded. The CSV is the only format available in analyzed data sources that is an Internet mime-type, an open file format (Shafranovich, 2005), an Internet tabular data model (Tennison, Kellogg and Herman 2015), and machine-readable by all well-known programming languages (Lebo and Williams 2010). Moreover, CSV is the only format that appears as an export option in all the interfaces analyzed.

Data analysis

In order to systematize the data analysis, concepts from Entity-Relationship (ER) Modeling (Silberschatz, Korth and Sudarshan 2011) were applied; using the set of conventions from ER "[…] to assist in databases design processes" (Date 2016, p. 64).

An ER model was developed (Figure 1), designed to store data collected from (i) services, (ii) resources available in each service, (iii) datasets available in each resource, and (iv) attributes available in each dataset.

In addition, two tables were developed to store information about controlled vocabularies applied in datasets and attributes, in order to control the set of available formats and data types in these elements (Date 2016, p. 228).

Figure 1: Diagram of Entity-Relationship model developed for data collecting
As a first step, the ER structure was applied in an Open Document Spreadsheet file (ODS) to store data collected in this study, which was able to identify in the 82 datasets a total of 2,280 attributes, with a subset of 1,342 distinct attribute labels.

The second step was to convert the ODS file to a CSV file and upload and import it into a database management system (DBMS) for subsequent data analysis, with tables and columns representing ER model entities and attributes, respectively.

A script for Python programming language was developed to assist the processing and reordering of data uploaded to tables in the DBMS into a second data structure, adopting the online analytical processing cube as a three-dimensional representation of services (s), datasets entities (e), and attributes (a), acting as perspectives of analysis (Gray, Bosworth, Lyaman and Pirahesh 1996; Inmon 2005; Kimball and Ross 2011).

The collected data was reordered to OLAP cube dimensions by concepts derived from the pivot table process (Cornell 2005).

To evaluate an OLAP fact, we intended to observe the intersections of the OLAP cube to determine the characteristics shared internally and externally by services, entities and attributes that may affect semantic issues of data collection.

Results

The data analysis identified several attributes with label names composed by filtering, grouping or sorting specifications as a part of text, a pattern followed only by online social network statistical data export tools. This leads to an increase of complexity involved in how to interpret those attributes and label inherent characteristics as values, using fully or semi-automated data collecting algorithms.

In this scenario, it is possible to determine that an entity (e) may have two attributes (a and a) sharing the same semantics (S), even when both attributes show distinct text labels in data collecting, expressed by the formula:

\[ S(e_x, a_x) = S(e_x, a_y) \]

An example that fits in this formula are two attributes (a and a) from a unique entity (e), with filtering specifications as a part of text labels, representing the total of people who engaged with the journal content on an online social network profile (S) by geographic area.

From another perspective, results from data analysis identified a larger set of attributes that do not relate to any description of its content, formed by 88.69% of available attributes, which means that label is the only explicit information on those attributes available at the moment of data collection.

---

6 The data collected in this research are available at http://dadosabertos.info/data/collection_recodaf_2017.

7 Also known as OLAP cube.
Furthermore, all attributes that share equal label names are part of a subset of attributes that do not have any description. This is critical in a data collecting process, primarily because this subset of attributes plays a significant role in the interoperability of data, inherently capable of being part of the set of potential primary keys with unique value restrictions, helping to build relationships between data sources, or determining geographic, temporal or linguistic aspects of the content itself.

This absence of semantics, with the exception of the availability of text labels, does not ensure that attributes of two distinct entities ($e_x$ and $e_y$) that share equal labels ($a_x$) will, consequently, share the same formal semantics ($S$) in data collection by external agents, expressed by the formula:

$$S(e_x, a_x) \neq S(e_y, a_x)$$

That effect requires external teams to interpret the semantics of these elements locally, aided by their skills or previous knowledge.

For example, two attributes that share equal text labels ($a_x$) from distinct entities ($b_x$ and $b_y$), without proper description of their content, may require interpretation of formats, data types, primary keys, unique restrictions, and controlled vocabularies applied, increasing the risk of wrong interpretations of values, thus preventing data collection teams from understanding that attributes may share the same text labels but not the same semantics ($S$).

**Conclusions**

Data analysis helped to identify the critical points related to the adherence of descriptive elements in the datasets analyzed, especially the lack of descriptive elements in the data collection process when triggered through the available export interfaces.

To reduce this dissonance between attributes, export interfaces could provide more semantic information bound to datasets. This information may be fundamental to interpret data available from different sources. Therefore, one action to reduce semantic dissonances between attributes is the enhancement of text labeling rules, including the use of controlled vocabularies and restriction clauses.

Moreover, the semantic dissonances in these entities may interfere with the development process of relationships between attributes from different datasets, thereby reducing the potential for interoperability.
References


Luciana de Souza Gracioso, Letícia Reis da Silveira, Maria da Graça Simões, Luzia Sigoli Fernandes Costa

Image organization on the Web: an analysis from the perspective of cultural heritage of rural farms in Brazil

Abstract
The main objective of this work is to examine the organization methods of images in the field of information science and verify if the current imagery-related information systems available on the Web put into practice the guidelines on processing thematic photographs found in the scientific literature in the area. Specifically, the goal was to create, from such methodologies, a tool to analyze some of these systems, termed an “Observation and Analysis Guideline”. The work aimed to characterize the selected information systems, by applying the above guideline in the search and retrieval user interfaces, performing an analysis from the user’s perspective, and thus obtaining new parameters that were possibly not suggested in the methodologies studied. The specific study objective also includes applying the guideline to Virtual Memory software in order to validate the categories developed for the description of cultural heritage aspects (albeit focused on photos), verifying that they meet the best practices suggested by the literature and, if necessary, provide parameters that can assist in the processing and retrieval of images by the system.

Introduction
The study identifies to what degree certain online systems available for organizing and retrieving photos include some of the main categories related to thematic processing of photographs in their user search interface, as researched and reported in the literature of Brazilian Information Science, but which were developed from an international methodological framework. Among these systems, special attention was given to the analysis of Virtual Memory, a software developed within the Project “Criteria and Methodologies to perform the São Paulo Cultural Rural Heritage inventory”, financed by FAPESP, Federal University of São Carlos, Brazil, which proposes the construction of a standard description of information covering specific fields for the processes of indexing cultural heritage images. The objective of the Virtual Memory construction project is to protect and make available all types of knowledge related to historic farms, including photographs. In this context, parallel to the study of image organization systems, this research provides assistance for detailed guidance with regard to indexing historic photographs of Brazilian rural heritage. A photograph registration simulation was performed using the Virtual Memory software in order to define whether it includes guidelines for organizing photographs defined by information science, and thus validate it as an organization and retrieval information tool for this field. The main objective of this work is: (i) to create a tool, termed an “Observation and Analysis Guideline”, using image organization methods from the field of information science, in order to analyze some online image information and
retrieval systems; (ii) to characterize the selected information systems, applying the above guideline in the search and retrieval user interfaces, performing an analysis from the user’s perspective, and thus obtaining new parameters that were possibly not suggested in the methodologies studied; (iii) to apply the guideline to Virtual Memory software in order to validate the categories developed for the description of cultural heritage aspects (albeit focused on photos), verifying that they meet the best practices suggested by the literature and, if necessary, to provide parameters that can assist in the processing and retrieval of images by the system.

**Methodology**

This work is defined as exploratory research because in studying images, it seeks to analyze and define the best practices made available to users regarding the processing of this type of information system, while identifying such practices in the systems selected. Besides being exploratory research, it can also be defined as intervention research, because as previously mentioned, one of its objectives regards the analysis and suggestion of possible parameters for enhancing Virtual Memory. Methodologically, this is a bibliographic and documentary study of an exploratory nature, partly developed as intervention research and using participant observation to collect data. Following this introduction, we present the detailed methodological approach that was undertaken, considering the observation feedback of the guideline developed for systems analysis. We also consider the research outlined in the literature aimed at retrieving scientific knowledge on a particular topic in national and international databases, specifically for organization of information focused on images. Since the main aim was to study user interfaces of Web systems, we conducted participant observation, and given that these interfaces can be considered as documents, our work can thus also be considered as documentary research. It should be emphasized that this study was cross-sectioned, since the selection of the systems took place at a given point of the research.

It should also be clarified that to apply the observation guideline, the analysis was condensed into three main scopes: the academic scope (systems developed by higher education institutions), the institutional scope (systems developed by institutions dealing with cultural heritage information), and the social scope (systems developed for photo registration in Web environments). These institutions were all analyzed at global, national and local levels. Flowchart 1, shown below, was prepared to simplify its understanding and summarize the phases covered during this work.

The flowchart essentially explains the study of information science approaches to analyzing and organizing images, and thereafter parameters were set for preparing an observation and analysis guideline of previously selected systems. After constructing the guideline, it was applied in the selected systems, identifying new parameters that were later incorporated into the instrument. The guideline was then rewritten in order
Methods, techniques and challenges for the organization and indexing of photos

The photograph as a document is also a source of information, so that in addition to getting information from it, it is a fundamental object to assist people in forming knowledge, helping them to visualize, assimilate and understand certain concepts. For photographs to effectively convey information, they need to undergo a processing, which besides allowing them to be retrieved for use, also adds knowledge for the user. Regarding this thematic processing, its classification and indexing is outlined below.

When we address thematic processing of a photograph, it is necessary to take into account the different types of treatments from the various types of units that address such informational content, whether it be archives, museums, memorial centers, personal files, libraries or the Web itself. Each of these information units adopts a suitable methodology for processing its materials. Therefore, the objective is to address some of the image analysis techniques in the literature of information science, essentially focusing on the national literature, partly as a means of highlighting the progress made in this regard in Brazil and to encourage further studies, but also because we are analyzing a national scenario for processing information about historic farms.

Below there is a summary of the various methodologies identified, highlighting their key aspects concerning the representation and organization of image content. This served as a support for preparing the guideline tool for observing and analyzing user interfaces of systems that process Web photographs:

- MANUAL OF THE NATIONAL LIBRARY (of Brazil 1998): One of the first initiatives for representing images compiled by librarians from the National Library and used until the present day by professionals from various fields. It takes into account the historical aspect of illustrative materials, asking: Who photographed? When? Where? What and/or who was photographed? It also proposes that the image must pass through the librarian or historian to give a short summary and from this summary select the descriptive terms for later retrieval, always moving from broader to more specific terms. Use of controlled vocabulary and natural language.

MAIMONE (2007): This approach covers mainly pictorial contents of an artistic and pictorial nature, but can also provide parameters for indexing other types of documents, such as photographs. The author relied primarily on the methodology developed by Agustín Lacruz (2006) and adapted it within a Brazilian context, in order to standardize the indexing of documents, with regard to the essential elements for effectively representing information and adding to user knowledge about museums and art galleries. The importance of this study thus lies in the measures it provides for processing materials in museums, especially pictorial information. The summary field is highlighted in this work, as well as the use of natural and controlled language for indexing.

TOREZAN (2007): In this study, the author strives to approach photography as a source of information and understand the analysis of the photographic document within its visual and historical possibilities. It uses questions such as “Which elements are relevant in the description of the photographic document?” And “How should the relevant information of the photograph be organized?” to guide her work. Therefore, steps are set out to describe the image document, namely: Primary Analysis, Secondary Analysis, Supplementary Analysis, Support Analysis, Technical Production Analysis and Historical Research. These steps, according to the author, are sufficient to generate a new document about the photograph, which in turn constitutes a critical element for image indexing.

COSTA (2008): Perhaps the most complete methodology for the analysis of photographs. Through a set of procedures from information science, literary and rhetoric narrative, this study formulates a model for the analysis and representation of historically-based advertising images. It uses content analysis of rhetorical discourse, based on concepts and terms related to Ranganathan’s essential categories (personality, matter, energy, space and time), extended to include each literary narrative category. Associated to this, the author also proposes that the image must be analyzed considering its denotative aspects – captured at a first-level of analysis, with visible and conceptually explicit signs – as well as its connotative aspects – noticeable at a second-level of analysis, with signs read as values, emotions and attitudes – thus taking into account all the general and specific historical context in which the image was produced. The rhetorical discourse analysis helps professionals to identify both denotative as well as connotative aspects of the image and its various meanings, which according to Costa (2008) can generate a differentiated documentary product, with higher added value. Finally, in literary theory, the historical and sociocultural context in which the work was produced is sought, contributing to the analysis of image content. It advocates the use of natural language for
indexing images.

• MARTINEZ (2009): This study focused on photographs as the study object of three areas of information science – museology, biblioteconomy and archiving – analyzing the application of the Brazilian Standard for Archival Description (NOBRADE) to photograph description in these areas, enabling the exchange of information between different types of institutions. As a result of the comparison, the author prepared a table to simplify visualization of their similarities and differences, to allow clear identification of similar or non-similar elements. According to Martinez (2009), it can be concluded that NOBRADE encompasses all the elements in the research instruments of the institutions analyzed, and as such this standard can play an important role in the photographic description process, fulfilling desired goals.

From the guidance provided by these approaches, the most commonly processed categories were extracted, which resulted in obtaining parameters for how to include elements in the guideline. This tool was divided into three parts, in order to evaluate aspects relating to content analysis, information representation and information retrieval in the systems selected.

Results and final remarks

As a result, a guideline tool was developed to diagnose image representation practices, which was later validated and used to evaluate Virtual Memory. Table 1 summarizes the results of applying the guideline in the systems chosen for analysis, now including the results from its application to Virtual Memory. The categories are separated according to the three dimensions of analysis, using “S” and “N” which mean “Yes” and “No”, respectively:
Table 1: Guidelines’ application results in the systems

<table>
<thead>
<tr>
<th>Allows free Registration of photos?</th>
<th>Corbis</th>
<th>Fund. Casa Rui</th>
<th>Barbosa</th>
<th>SACI</th>
<th>VIA</th>
<th>Dedalus</th>
<th>Fundo Florestan</th>
<th>Ferrandes</th>
<th>Flickr</th>
<th>Memória Virtual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the image contextualized (within a collection)?</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Is there a file for informative description of the image (title, date, place, support, dimensions)?</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Is there a field for describing the size of the image?</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Is there a field for summarizing the content of the image?</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Is there a field for the historical context of the image?</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Is there information about access conditions (existence and location of original documents, existence and location of copies, etc.)?</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Are the images indexed by subject and/or keywords?</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Is there a field for social indexing (subject responsible for tags, groups of photographs corresponding labels)?</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Is there a field for image caption?</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Is controlled language used for indexing?</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Are there search tools?</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Can the user perform advanced/integrated search?</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Is the image retrieved with good resolution?</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
</tbody>
</table>

The guideline prepared was applied in seven websystems that processed image content, in order to determine which categories have been developed and used by these systems for analyzing and organizing photographs, and also to verify whether these current systems include the techniques suggested by the scientific literature in the area of information science with regard to the organization and representation of information, as well as aspects regarding their search tools.

The results showed that that systems take into account most of the photograph analysis practices suggested by the methodologies studied; they provide users with
advanced research tools in their interfaces, often innovating the categories to refine the search tools, such as the Corbis system that uses color filters, number of individuals in the photograph, similar photos, and other factors; they are concerned about providing the user with good information available on the interface and also a user-friendly system. Furthermore, the implementation of the guideline allowed us to diagnose a growing concern of the systems regarding social indexing, and four of the seven major systems examined had some type of access to this type of indexing – Corbis, Casa de Rui Barbosa Foundation, USP Integrated Search Portal and Flickr.

A second objective of this study was to examine Virtual Memory software, developed to group, organize and provide access to information. The specific focus was on how it dealt with Brazilian historical heritage in the context of 19th century coffee production farms, assessing whether this system also took into account the practices suggested by the scientific literature in the area regarding analysis and description of photographs, and to evaluate it in respect to content retrieval. Applying the guideline tool to analyze Virtual Memory software thus enabled us to conclude that it is a very complete system which broadly explores all the levels needed to thoroughly describe cultural heritage, including photographs, positively responding to the guideline parameters prepared for this work, with some exceptions, such as the expressive dimension field and social indexing.

In the case of Virtual Memory, aspects regarding expressive dimension can also be incorporated in the index field for the summary of the photograph. A further suggestion to optimize the preparation of the summary and to include expressive dimension aspects could be to incorporate a “template”, or guide, to direct the indexer when preparing the summary, for example, including fields with the expressions “Who? What? Where? How? When?”, as suggested by IC methodologies, to be filled in and also to guide the indexer to observe and identify issues related to the expressive dimension of the photograph. These could be ways to improve the quality of the summary, making it richer more detailed.

This research also argues that social indexing should be taken into account, because the user’s socio-cognitive perspective when indexing or searching for information can, in fact, contribute greatly to a more efficient retrieval of search results. One suggestion is to indirectly use the users’ perspective, as their search results could help to index terms. If the system is capable of storing the search terms used by users, these terms could be retrieved and subsequently analyzed by a professional cataloger, or reviewer, and added to the fields of topics and descriptors of the cultural heritage they relate to.

Another field to be discussed in relation to Virtual Memory, which the guideline analyzed, regards the system’s search tools. It is known that this system is still in development, and that it may be the most complete means of processing cultural heritage information. It is therefore crucial that it contain search tools that optimize the
users’ results. Through analysis of Web systems, the main trends regarding their search tools were observed, showing that with the exception of SACI, all systems had advanced search tools able to integrate their search terms through Boolean operators, and search by different categories, different databases, etc.

Thus, regarding the Virtual Memory proposal, we consider that an advanced search tool should be added, one that allows users to filter their searches by material types (Archaeological, Archival, Audiovisual, Bibliographic, Built, Mobile and Integrated or Landscape, according to the system’s categories) and also by institutions that have registered their cultural heritage assets, for a more specific search.

Besides the traditional fields available for advanced searches, such as authors, titles, locations, dates and subjects, the analysis of the Corbis system interface showed a multitude of search types that can further optimize the results for photographs in the system, such as filters that determine surveys by the color of the photos, number of people in the photo, gender, age, ethnicity, types of images (photographs, illustrations...), image orientation (horizontal, panoramic, square, vertical...), style (outdoors, indoors, silhouette...), visualization (head and shoulders, full body, empty space, cropped image...), panorama (aerial view, bottom view, top view, looking at the camera, looking away from the camera...). Other ways of further refining search results include looking for photographs similar to that found, or using keywords employed for indexing the image as a source for new searches. These are some of the interesting examples found throughout the study and which could enhance the search for historic photographs using Virtual Memory software, through the evaluation and definition of its needs. Essentially, these are just a few things to be considered, in addition to assessing their feasibility within the system.

As with most studies, the current work also encountered some limitations and methodological challenges, such as the difficulty in building research instruments and then defining the methodologies used, especially as this is a research-intervention study. Another difficulty regards the number of systems and photographs analyzed. Nowadays, many systems are developed for processing photographs, and there are thousands of pictures indexed by them. However, the relatively short time available for the research did not allow us to conduct a more in-depth analysis, analyzing more systems and more photographs. It is thus important to enable this research to continue and with it the development of new studies related to the theme, for example, applying the guideline tool in more systems in different contexts, to test their effectiveness.

Finally, as presented here, there are many forms of information, be they text, image, audio or sound, which need to be disseminated so that they make sense and can help form users’ knowledge for effective decision making, completion of informational gaps and also for the retrieval and preservation of memory. Image information, in turn, has increasingly gained ground as a source of information, transmitting knowledge as a
whole or to supplement other information sources. As demonstrated in this work, the increasing concern of institutions to form collections and make them available to the public proves this fact. Thus, it is vitally important to take into account the best practices for the processing of this information so that it can fully and satisfactorily reach the user.

With the efforts made in this work and the construction of a diagnostic tool based on observation and analysis, we believe that it can be useful to other institutions and studies regarding decision-making processes and for assistance in relevant practices for analyzing and organizing pictures, especially with regard to historical images.

References
Information access in the digital era: document visualization strategy

Abstract
In this work, we focus on the document visualization strategy to support access to information in the digital era. First, we discuss the dynamics of the document visualization approach and the ability to generate innovations with a direct impact on the competitive digital transformation scenario. Second, we discuss the visualization and computational intelligence methods such as data mining and knowledge discovery as important tools to improve the decision-making process. We then present the concept of knowledge organization systems and the main challenges related to document visualization strategies. Finally, we discuss the increasing prevalence of digital and visual literacies in relation to how we read and view information and communicate with others to meet the demands of the changes brought by the digital age.

Introduction
Continuous improvements in technology, networking, software and hardware, including mobile and wireless services and protection of content and services, have made possible the development of advanced digital content. Greater cooperation is a major challenge, since the production of digital content requires agreements between content developers, equipment manufacturers and information distribution.

Digital information now occupies a key place in organizations’ decision process. Libraries are acquiring digital information and providing access to users. Most of the information accessed today is digital and the Internet is playing a decisive role in these processes. Even analogue information is being converted to digital for ease of access and use.

Digital content has become a major driver of change in this digital age. Technological innovation and current information users’ demands are leading to new and direct ways of addressing knowledge organization, new methods of access, use and appropriation of information. Research results, for example, are becoming more accessible, and digital content requires new tools for information searching.

Technological innovation and the ubiquity of communication tools, economic uncertainty, changes in workplace and educational structures, the global economy, generational differences, the blurred distinction between the production and consumption of information, and heightened national security production and consumption of information are just some of the factors affecting the creation of knowledge in the digital era.

An important part of a digital library is the ability to access stored information effectively. Due to recent developments in information retrieval, a digital library is usually equipped with an automatic search and retrieval system which library users may
employ to find documents.

Against such a backdrop, this article proposes to examine a document visualization strategy to support information access in the digital era. Regarding research methodology, we will implement a case study research method, focusing on the necessary management changes to put such a strategy into practice within the workplace. Scrutiny is hereby placed on each key stage of the process, not necessarily the long-term outcome.

**Document visualization approach**

Document visualization is a class of information visualization techniques that transforms textual information such as words, sentences, documents, and their relationships into a visual form, enabling users to better understand textual documents and to lessen their mental workload when faced with a substantial quantity of available textual documents. Document visualization has significant benefits in helping users to analyze and control large amounts of textual information (Gan 2013), as shown in Figure 1.

![Figure 1: Document Visualization – Function Observatory Information](https://www.adbs.fr/lobservatoire-de-fonction-information)
With the exploding amount of document information and sources, efficient and intuitive visualization tools are desperately needed to assist users in understanding the contents and features of a document, while discovering hidden information (Gan 2013).

Data visualization is an extensive field at the crossroads of mathematics, computer science, cognitive and perception science, engineering, and physics. The goal of data visualization is to use images to improve our understanding of a dataset, drawing on techniques from mathematics, computer science, cognitive and perception science, and physics (Telea 2007).

A huge amount of data on the Web has geographic features. The World Wide Web Consortium (W3C) and the Open Geospatial Consortium (OGC) want to facilitate the Web of Data development using the Web infrastructure as well as the infrastructure of geospatial systems, as shown in Figure 5.

Figure 2: Web of Data and Geolocation Development

One of the greatest scientific challenges of the 21st century is how to master, organize and extract useful knowledge from the overwhelming flow of information
made available by today's data acquisition systems and computing resources. Visualization is the prime means of taking up this challenge.

Visualization is the process of representing data, information, and knowledge in a visual form to support the tasks of exploration, confirmation, presentation, and understanding. Scientific visualization is concerned with techniques that allow scientists and engineers to extract knowledge from the results of simulations and computations. Tremendously large collections of numerical values, which contain a great deal of information, are being produced and collected. The problem is to convey all this information to the scientist so that effective use can be made of it. This requires a method of communication with a high bandwidth and an effective interface. Computer generated images and human vision mediated by the principles of perceptual psychology are the means used in scientific visualization to achieve such communication (Bonneau 2006), Figure 6.

Figure 3: Scientific visualization

Source: NASA/Goddard Space Flight Center Scientific Visualization Studio
To accomplish all these goals as far as visualization technologies are concerned, the use of digital technology is evolving toward comprehensive solutions to manage information systems using a single repository and a single interface, dramatically reducing the costs and complexity of managing technological resources, including mobiles, IoT - Internet of Things, tablets, and networking devices. Innovations continue to emerge at a frenzied pace, driven by the rapid advancement of technology and the big data phenomenon. Information is a key asset in the digital era, pressing the organizational environment to produce tangible and sustainable results (Paletta 2008).

Despite the importance of the connection between technology, innovation and economic prosperity, organizations must be able to adapt and evolve if they wish to survive. Businesses operate with the knowledge that their competitors will inevitably come to the market with a product that will change the basis of competition.

According to Trott (2010), the ability to change and adapt is essential to survival and the idea of innovation is widely accepted. It has become part of our culture. In the competitive digital transformation scenario, visualization and computational intelligence methods such as data mining and knowledge discovery are important tools to improve the decision-making process.

Knowledge Organization Systems – the IAEA case experience

In the context widely known as the world’s “Atoms for Peace” organization within the United Nations family, the IAEA is the international centre for cooperation in the nuclear field. The Agency works with its member States and multiple partners worldwide to promote the safe, secure and peaceful use of nuclear technologies. According to the IAEA (2015):
For managing knowledge organization systems which follow W3C standards, a tool has been acquired which allows users to create KOSs, transform existing ones into the SKOS format and import them, edit KOSs, and publish them on several media, most prominently the Web. In addition to the basic functionality of managing KOSs, the tool offers features for interlinking KOS concepts, and linking the concepts to external data. This opens onto the world of linked data, the Internet of data in which data are treated in an analogous way to documents. Obeying the naming conventions of each single term by URIs (unified resource identifiers), data may be interlinked with other data, forming a knowledge network. The number of information sources which are published in this way grows continuously as shown in Figure 2 (AIEA 2015).

[…] this allows the construction of extensive knowledge graphs to enable discovery of knowledge, meaning that a search will return not only results by full text search, but also related items. For representing knowledge domains, this technology, often referred to as “semantic technology”, allows the generation of content-rich knowledge models, providing the user with a navigable knowledge map which may be queried with a specific query language (SPARQL), and which are often supported by visualization and refined semantic search tools (AIEA, 2015).
Semantic technologies are not specific to an application area or knowledge domain. The distinguishing feature of semantic technologies instead consists in building applications based on knowledge organization systems and/or taking advantage of linked (open) data sources.

An essential point to note is that if an organization is to be fully effective, every part of that organization needs to actively contribute to innovation (Goffin and Mitchell, 2010). Innovation should not only originate in the R&D department in a manufacturing company or the strategic planning group in a service operation. The functional areas that should be involved are:

- Research and Development
- Marketing
- Operations
- Finance and Accounting
- Human Resource Management
- Outside Resources

According to AIEA (2015), knowledge organization systems provide powerful means of modelling knowledge domains. Linked data allows a KOS base to be enriched with other sources which are increasingly being published by many organizations on the web. Publicly available sources such as Wikipedia and restricted, topical data sources may contribute significantly to enhance established concepts, leading to rich knowledge bases which lend themselves to knowledge discovery as described previously. Such knowledge bases provide an excellent backbone for constructing knowledge portals, enabling improved search by automatically categorizing documents, translating their taxonomical hierarchy into site pages, and linking pages to internal and external information sources.

The more changes there are in the technology of products, services, and operations, the more changes take place in administrative procedures – new strategies, new organizational structures, and new operating processes will be required to successfully capture the potential benefits of the venture. The failures of technological change typically occur when either too much technology is adopted too quickly, or not enough technology is adopted to stay ahead of the competition (Ettlie 2006).

**Conclusion**

We need to have both digital and visual literacy skills to evaluate, use, and create digital resources. Being able to read screens with deeper understanding and to interpret meaning from pictures enables us to make educated judgments about what we find online. Shaw (2016) states that digital literacy refers to the competency to:

- Read and Write with technology tools
- Find, evaluate, utilize, share, and create content using information technologies
and the Internet
- Use and manage technology proficiently
- Comply with ethical and legal technology use

Visual literacy is the ability to draw information and inferences from photographs, pictures, or illustrations, whether still or moving.

[...] visual literacy is a set of abilities that enables and individual to effectively find, interpret, evaluate, use, and create images and visual media. Visual literacy skills equip a learner to understand and analyse the contextual, cultural, ethical, aesthetic, intellectual, and technical components involved in the production and use of visual materials. A visually literate individual is both a critical consumer of visual media and a competent contributor to a body of shared knowledge (Shaw 2016).

Accordingly, Pratish (2006) states that visual literacy refers to the competency to:
- understand the subject matter of images,
- analyze the syntax of images including style and composition,
- analyze the techniques used to produce the image,
- evaluate the aesthetic merit of the work,
- evaluate the merit of the work in terms of purpose and audience, and,
- grasp the synergy, interaction, innovation, affective impact and/or ‘feel’ of an image
- analyze and interpret images to gain meaning within the cultural context the image was created and exists

Digital and visual literacies have become common to how we read and view information and communicate with others. Digital and visual literacies help us interpret what we experience, analyze what we are exposed to, and make conclusions using our critical thinking skills (Shaw 2016)

In order to further research in information science, it is essential to continue studying technology transformation with regard to how users search, access and use information in the digital era and the role of document visualization strategies to help improve new knowledge development in the competitive global economy.


References


Abstract
During the construction of catalogs, many requirements are not covered and are unavailable for informational
treatment. For this reason, instant information and innovation centers need to consider how to improve their
user environments, as well as enhancing informational treatment of their information resources, using
semantic technologies and Linked Data principles. Thus, the objective of this paper is to present international
and Brazilian initiatives that seek to integrate libraries, archives, museums and art galleries in relation to the
descriptive content of each documentary context. The research is characterized as qualitative and exploratory,
adopting a bibliographical survey method. The results describe Brazilian and international initiatives.
Accordingly, we discuss the crucial role played by Linked Data in the dissemination of information for users
and consider its use as a tool to improve the sustainability of data in information centers, systems of
dissemination and organization of information resources.

Introduction
The strategic use of technologies in libraries, archives, museums and art galleries
can advance various products and services, offering users a better experience with
access to information. Technologies can also be used to broaden the possibilities of
accessing, locating and retrieving resources in information systems, such as catalogs
and research tools.

Informational treatment involves processes and procedures using control and
standardization instruments that make it possible to present an informational resource
in such a way that its representation alone identifies it as unique and refers to the
properties of the original. Accordingly, a fundamental basis is attributed which also
provides clarification for cultural institutions in how to construct catalogs and research
instruments, given their need to transpose properties into elements of identification and
access for each informational item.

However, during the construction of catalogs, many requirements are not taken into
account, generating difficulties during informational treatment, such as duplicity of
data, which can potentially affect the recovery of these resources by users. The process
of conceptual modeling of data is highlighted as a possible solution, but new
technologies from the semantic Web have further enhanced user experience with
cataloging, as already addressed in research by Simionato, Arakaki and Santos (2017)
and Arakaki, Simionato and Santos (2017).

For this reason, information centers need to seek innovations and consider creating
an improved environment for users, in addition to focusing on the treatment of their
information resources. Thus, the objective of this paper is to present international and Brazilian initiatives aimed at integrating libraries, archives, museums and art galleries in relation to the descriptive content of each documentary context.

**Integration of archives, libraries and museums by descriptive treatment**

As reflected in the literature, descriptive treatment performed in archives, libraries and museums is convergent in many respects, yet there are distinct differences in each case. The informational items in libraries and museums are recorded individually, but in museums, contrary to the procedure used in archives, each document corresponds to a set of documents, referred to as its documentary background.

In short, the descriptive process in archives involves three theoretical currents that guide the management of documents; namely, integrated, diplomatic and post-modern archiving. Among them, of particular relevance here is the integrated archival system proposed by Rousseau and Couture (1998) for guiding description from the current phase of a file, “[...] construction of efficient mechanisms for access, relationships between the documents themselves and their management activities and control of the acquis in all their ages” (Simionato 2015, p. 83). In libraries, cataloguing aims to construct the forms of representation to feed catalogs from the standardized description of informative resources, contemplating their form, content and arrangement in collections (Santos 2013). Museological description is a process very similar to cataloging, whereby records are created to dictate the organization of a catalog and establish control over the organizational aspects of all the informational resources in the institution (Ferrez 1994). Art galleries are more commercially oriented, although their means of representation is very similar to museums, determining how data from these environments is reused.

With such diversity of patterns and schemes for resource descriptions, there is clearly a tendency from several institutions to collaborate in building structures that can exchange information, even using heterogeneous metadata patterns. Marcondes (2016) points out that the most trivial step in the harmonization of contexts is to overcome the isolation of existing systems in archives, libraries and museums, so as to integrate them into other more complex systems existing on the Web.

In this way, Carrasco, Thaller and Vidotti (2015) reveal that among the challenges in the process of harmonizing the metadata of different institutions, there are specific difficulties with structures such as those of data models that are designed according to the requirements of each community, without always considering the interoperability requirements between them. The interoperability and planning of catalog construction for libraries, archives, museums and art galleries should be one of the principles for supporting information resource records. This highlights the importance of conceptual modeling so that records are increasingly relatable, avoiding monolithic records with
duplication of data.

Thus, interoperability and data reuse are key to minimizing the efforts of describing informational resources in heterogeneous environments, but for a system to be able to accomplish this, it must have the necessary set of tools, technologies and structures.

Among one of the key aspects for interoperability, is the linked data that is characterized by the principles and best practices for publishing and linking structured data to other datasets, to facilitate the search for human and non-human agents. Linked data are thus principles that guide how to make structured data bindings using protocols, Uniform Resource Identifiers (URIs), and the Resource Description Framework (RDF). In the proposed use of open data, open formats and licenses, the term used is linked open data (LOD).

**Methods**

This work is characterized by qualitative and exploratory research, as it seeks to familiarize and analyze the linked data initiatives between libraries, archives, museums and art galleries and their development. For the localization of these documents (dissertations, theses, articles, events, scientific reports and books) the following databases were used: Reference database of journal articles in Information Science (BRAPCI), CAPES journals, Scientific Electronic Library Online (SciELO), Web of Science, in the Brazilian Digital Library of Theses and Dissertations (BDTD). This study selected those works that address the issues of integrating description within informational units, in addition to texts deemed relevant to clarify this work.

**Main results**

Concern about cultural heritage data is growing, especially with the processes of publishing, linking and curating such data (Kirchhoff, Schweibenz and Sieglerschmidt 2009; Novia 2012; Hooland and Verborgh 2014). As a result, one key initiative is Linked Open Data in Libraries, Archives and Museums (LODLAM) because it involves three types of institutions with the possibilities of open data binding practices for linked open data. The LODLAM movement was initiated in 2011 as an event with four editions thus far, according to Voss (2012).

Moreover, the purpose of bringing together the conceptual structures of libraries, archives and museums is reiterated by Simionato (2015) with the conceptual model of digital images for libraries, archives and museums, the DILAM model. Another initiative is the harmonization of FRBR with CIDOC CRM, which consists of promoting interoperability of cultural heritage information with the objective of “[...]
provide common language for heterogeneous information systems, and enable integration, despite possible semantic and structural incompatibilities.” (Carrasco, Thaller and Vidotti 2015, p. 212).
In addition to LODLAM, there is a movement to also incorporate art galleries as is the case with Linked Open Data Galleries, Libraries, Archives and Museum (LODGLAM). Similarly, with the aim of providing and preserving European cultural heritage, Europeana is a digital library that gathers several available collections from various European countries. To standardize and create a minimum structure of description for these resources, a model called Europeana Data Model (EDM) was created. This initiative is also related to OPENGLM – Galleries, Libraries, Archives and Museums, which promotes the use and open access of cultural heritage. OPENGLM is an initiative founded by the European Commission and coordinated by Open Knowledge (Arakaki 2016).

In the American context, cooperation efforts between libraries and museums include Digital Public Library of America (DPLA), a digital library that aggregates metadata for various types of informational resources around the United States and is used as the basis of its application profile DPLA Metadata Application Profile and the EDM. This reinforces the importance of Europeana in the process of integrating informational resources into various systems.

Europeana’s prominence on the world stage is shown by the fact that it currently accounts for over fifty million cultural objects from libraries, archives and museums, which can be accessed both through the project’s portal and through SPARQL Endpoint, allowing structured publication of this data and for it to be linked with other datasets. However, the scope and complexity of Europeana in gathering data from different countries led to the creation of a data model itself, the Europeana Data Model (EDM), which is used in conjunction with other standards and vocabulary of the W3C Consortium.

Europeana groups together other projects from European countries, for example, the Culturaitalia project, which consists of a portal for Italian culture. This project aggregates cultural heritage data from institutions at national, regional and local level, and is directly linked to Italy’s Ministry of Cultural patrimony, activities and tourism (MiBACT) (Di Giorgio 2014). Another project is Linked Jazz, which is focused on the historical development of the jazz musical genre, using interview archives (oral history records) to establish personal and professional relationships between jazz musicians (Pattuelli, Miller, Lange, Fitzell and Li-Madeo 2013). This latest initiative is part of the broader list of linked open data movements applied in specific domains within libraries, museums or, in this case, files. In the same context, Pattuelli, Miller, Lange, Fitzell and Li-Madeo (2013) present projects from the Amsterdam Museum, the Smithsonian Museum of American Art, both focused on the museum domain, and the Social networks and Archival Context (donate), which deals with archives.

On the American continent, we highlight the initiatives Digital Public Library of America (DPLA) and Linked Jazz. The first is a digital library that aggregates resources
from libraries, archives and museums in the United States, in a similar way to the Europeana project, and its published data is aimed at the reuse and development of applications (Jones and Seikel 2016). Like the European project, the DPLA’s collection has references to cultural heritage, so that it can also be a partner project in national initiatives. Regarding libraries, WorldCat stands out, an initiative created by the Online Computer Library Center (OCLC) with the aim of making a bibliographic catalog linked worldwide. OCLC is also responsible for other initiatives, such as publication of the Dewey Decimal Classification (CDD) as linked data and participation in the Consortium Virtual International Authority File (VIAF). In addition, there are a number of services focused on linked data principles as well as those of the Getty Research Institute, whose vocabularies, presented above, are semantically structured for use in publishing and data binding.

On the Asian continent, the stand-out project is 'K-Culture time machine' whose objectives are to relate different cultural heritage databases through a metadata scheme, called Korean Cultural Heritage Data Model (KCHDM), and to develop an ontology of national heritage (Kim et al. 2015).

As presented here, there are numerous initiatives for the publication and linking of cultural heritage data, although the situation in Brazil does not seem to be following the trend. As a first step, the Brazilian Ministry of Culture (MinC) has announced its open data plan, which will provide, in stages, administrative and bureaucratic data from its secretariats, linked to organs and support systems providing information on regulations, maps, and centers pertaining to culture, arts and sports. These national movements relate to the Open Data Applied principles with regard to government data, a global tendency to encourage public transparency and which has led to information becoming accessible by law in Brazil. However, given the number of international initiatives, it is noted that open access to the national cultural heritage objects held in public institutions is lagging well behind the global tendency to publish open and linked data of such objects.

Conclusions

Therefore, with the increasingly intense flow of information, professionals need tools enabling the reuse of data to avoid the need for professionals to rework the representation of resources, as is happening in the United States and Europe. Consequently, the principles of linked data should be set up in proposals for cooperation between libraries, archives, museums and art galleries, which play a crucial role in the dissemination of information.

Finally, it is considered that efforts to integrate libraries, museum archives and art galleries are crucial to providing a connected environment and broadening users' navigation possibilities. The use of linked data principles thus enhances the discovery
of informational resources in catalogs, promoting access to new features in other digital information environments.

References


Vânia Mara Alves Lima, Cibele de Araújo Camargo Marques dos Santos

Knowledge Organization in a Web collaborative environment

Abstract
This paper describes the knowledge organization approach and challenges as the fundamental theoretical reference in the development of a digital iconographic collaborative Web environment that allows institutional users – such as GLAMs (Galleries, Libraries, Archives and Museums), NGOs, Universities and Research Centers – to collaborate with private users – such as students, teachers, photographers and the general public.

1. Introduction
According to Hjørland (2008), knowledge organization (KO), narrowly speaking, includes document description and indexing in libraries, archives, databases and other memory institutions. The field also studies the nature and quality of the knowledge organization process (KOP) and the knowledge organizing systems (KOS) that are used for the analysis and representation of subjects as concepts and controlled vocabulary.

The terminological standard ISO 1087 (International Standard Organization 2000) defines the concept as the unit of knowledge created by a unique combination of characteristics. Dahlberg (2014) highlights how concepts form the basis of any specific knowledge organization, leading to the creation of classes, and that the definitions of concepts are the resumé of content-determinant characteristics, the characteristic being the property of an object or a set of objects.

A controlled vocabulary is an indexing language whose purpose is to represent the topics of a particular area for information retrieval in the information system, whether physical or virtual (National Information Standard Organization 2010). Controlled vocabularies are designed for applications in which it is useful to identify each concept with one consistent label, for example when classifying documents, indexing them and/or searching them (International Standard Organization 2011). In the Encyclopedia of Knowledge Organization, Hjørland (2017) emphasizes that the indexing and retrieval criteria must take into consideration the subject fields and theories which characterize the domain-analytic approach.

The traditional approach of KO has been implicit in the management of databases with emphasis on the principle of controlled vocabulary, that is, by using a standard vocabulary to avoid synonyms and homonyms as indexing terms. The facet-analytical approach of KO refers to the semantic relations between terms and categories. In addition, the user-oriented views approach considers knowledge organization done by users as folksonomies. San Segundo (2013) refers to KO changes as a technological paradigm based on the process of interactive information search behavior, which has the effect of generating information feedback that flows in the digital context creating
folksonomies and retrieval tools based on the search.

This paper describes the challenges of using a knowledge organization approach as the fundamental theoretical reference in the development of a digital iconographic collaborative Web environment. This approach allows institutional users – such as GLAMs (Galleries, Libraries, Archives and Museums), NGOs, Universities and Research Centers to collaborate with private users such as students, teachers, photographers and the general public.

These challenges first involve conducting analytical-critical research on constructing both controlled vocabularies to index the iconographic information of an institutional collection, as well as the tags proposed by users for their private images, considering that it is necessary to standardize terminology for indexing and retrieval of information. Secondly, there is the process of indexing images, which can be understood as indexing the content of images themselves, as well as their textual content. Thirdly, it is necessary to find a solution to issues of metadata accordance between systems, the consolidation of standards which respond both to international interoperability requirements, and to local needs for organization and information access.

Our case study focuses on ARQUIGRAFIA, a nonprofit collaborative environment on the Web, with open public access, which is continually expanding and dedicated to visual culture and critical approaches to architectural images, with a special focus on Brazilian architecture and urban spaces. The main goal of this project is to bring together, within the same internet environment, institutions and private users committed to collaborating on research and collective knowledge with regard to architecture and urbanism in Brazil, as well as with Portuguese speaking communities around the world.

The main source of images used by ARQUIGRAFIA is the collection of the library of the School of Architecture and Urbanism of the University of São Paulo (FAUUSP), but the collaborative nature of the project sets it apart from institutional image databases on the internet, precisely because it involves a heterogeneous network of collaborators, both institutional and private, who jointly contribute to the construction and continuous growth of the digital image collection. The collaborators must fill in at least the title and author of the image, country and tags to upload an image to ARQUIGRAFIA, a procedure that can be performed on the website or the Android application for smartphones available for free download in Brazil at the Google Play Store. The system suggests terms selected from a list, but the collaborators can insert new tags in natural language.

Thus, ARQUIGRAFIA is an experimental project involving the active and collaborative participation of multiple users who are all engaged in the ongoing construction of the system. They are developing an "image constellation", with perpetual beta software, together constituting a Web community that constructs
knowledge collectively, respecting copyright, and using distinct and complementary interfaces such as a website and an Android application for smartphones.

2. Controlled vocabulary and architectural terminology

The fact that it is a collaborative environment means that ARQUIGRAFIA faces some problems pertaining to KO approaches: the principle of controlled vocabulary, the semantic relations between terms and categories, and the knowledge organization done by users. An example is the ambiguous meanings of terms such as “banco” – which in Portuguese means both “bank” and “bench” – generating distortion in information retrieval, as well as the tags included by users that do not have a terminological and vocabulary control.

Studies have demonstrated that the traditional principles for construction of controlled vocabulary would not be sufficient to solve these problems. It was necessary to develop a new model of controlled vocabulary based on delimiting the semantic field boundaries for each of the domain concepts designated by terms that represent the concepts in a specialized context.

In ISO 25964-1 (International Standard Organization 2011), the term is the word or phrase used to label a concept. Concepts, once defined, can be organized in a conceptual network, which supports logical and ontological relations between the terms in the controlled vocabulary. It is necessary to understand the concept as the unit of knowledge created by a unique combination of characteristics, and whose relationships with other concepts work as a particular system. The characteristic is the property of an object or a set of objects. The properties of the object are described in the definition of the concept that is designated by a term. Therefore, the definition of the concept must determine exactly what the limit of the term meaning is.

For terminology, the definition is a representation of a concept by a descriptive statement that serves to differentiate it from related concepts (International Standard Organization, 2000). This definition is the result of the compilation of true statements about the concept, that is, the definition states the essential and accidental characteristics about the object which the concept refers to (Dahlberg 1978). These statements make explicit the semantic relations that articulate the conceptual network, expressing the nature of each concept. As a result, they delimit the semantic field of the term, and must be operationalized in the controlled vocabulary.

The notion of semantic field is related to the work of Jost Trier in the 1930s, with regard to the German vocabulary of knowledge in the 13th and 14th centuries (Germain 1981). By bringing together the ideas of other authors within a coherent methodology, Trier shows that the lexical units in a language are organized in structured groups. Each unit is defined in relation to its position in respect to the others, that is to say, the meaning of a lexical unit is specified only by means of its similarity and difference in
relation to the other lexical units in the field.

This conjunctive and disjunctive process, which is used to put together the similarities and separate the differences found in the definitions of concepts and to identify the semantic field, is supported by the idea of significant traces. According to Hernando Cuadrado (1995), the minimal condition for words to belong to a common field is that they should have a common significant trace. The larger the number of significant traces, the more coherent the semantic field will be, and the smaller the number of words which are part of it. A word may be part of the semantic fields which are delimited by any of its significant traces. When a word has multiple meanings, each of them belong to a distinct semantic field. Thus, the semantic field of a word is made up of all the words that are related to it in the signified plane. For example, we can find historical building, abandoned building, commercial building, apartment building in the semantic field of the word building, because these words have the same significant traces in their concepts, which we can identify as the essential characteristics mentioned by Dahlberg (1978).

To equip controlled vocabulary for use in the collaborative environment ARQUIGRAFIA, we must not lose sight of its association with the ontological tools used in this context, which capture and explain the vocabulary used, mapping a given knowledge domain and, by indicating its general and specific categories, also act within the semantic field. Ontologies can be used as reference tools like vocabularies, providing a controlled and standardized terminology for indexing, but enriched with inference rules and axioms representing connections between different concepts. The construction of an ontology in the field of architecture was developed and tested in the system in order to improve image retrieval (Solis, Wasserman and Lima 2016).

The controlled vocabulary must also rely on domain terminology. Terminology, as a scientific discipline, develops rules for the construction of terminologies (as sets of terms from a specific domain), and for some time now they have been considered as the basis for controlled vocabulary construction.

The methodology of terminology for the construction of definitions seems today to be the most adequate way to serve as a basis for the mapping of a semantic field. Nowadays, the ARQUIGRAFIA team is undertaking the rigorous task of standardizing terminology among the lists of subjects used by the library for indexing the institutional collection of photographs and slides; the controlled vocabulary of the Integrated Library System of the University of Sao Paulo (VOCAUSP), the list of tags developed by the team, based on dictionaries and thesauri of architecture, and the tags used by the public to upload their private images in the collaborative repository. A survey was conducted on the keywords employed by users in order to gather new terms that could be defined and added to the controlled vocabulary. These tags belonging to the semantic universe of the users can enrich the vocabulary, show how users think and retrieve
The focus of this terminology standardization is the definition of terms. The terminological definition is regarded as an intermediate point between linguistic and ontological definitions; that is, a linguistic definition includes characteristics which allow one to discern one notion from the others in the same linguistic system, while an ontological definition consists of the exhaustive listing of the characteristics of a concept and is intended to include all possible information about this concept. Terminological definition aims at identifying a concept in the given subject context by listing the minimal essential characteristics that should be sufficient and necessary for this definition (Cabré 1993; De Besse 1997 in Azarian and Tebe 2011).

According to the terminological standard (International Standard Organization 2000), the definition of a term can be developed from the term’s extension or intension. The extension of a term consists of the group of all things which the term applies to. The intension of a term is the set of properties determining the objects to which the term applies. The intension is made up of the properties that a given object needs to have in order to be in the extension of the term. As an example, for the term *House*, we have:

- Term’s extension: *farmhouse; cottage; mansion.*
- Term’s intension: *a building made for people to live in; it can be made of wood, stone, bricks, concrete; it has different rooms.*

Therefore, an extensional definition explains a term based on the objects that it designates; by using the example given, we could explain the meaning of the term *house* by indicating a *farmhouse* or a *mansion*. The intensional definition explains a term by indicating its characteristics, in the example, *a building made for people to live in*, etc.

It seems clear that in order to achieve effective interoperability between tools for domain representation on the semantic web, the process should begin by mapping the several semantic fields that constitute the domain, or more specifically mapping the characteristics listed in the concept definitions that indicate the semantic field of the term. Next, we present the second challenge in the organization of knowledge in a collaborative environment such as ARQUIGRAFIA, namely the use of controlled vocabulary developed from texts on architecture in the process of indexing architectural images.

### 3. Indexing Architectural images

Architectural images are essential in architecture teaching and research, since an on-site visit to the world's major architectural works would be impossible, which emphasizes the need for a detailed and specific description of this material to serve its audience. These are records of places, buildings, cities, urban spaces, construction sites, and so on, forming the wide range of references required for the imaginary and visual
culture of architecture (Rozestraten 2010). A work of architecture can be better understood through the variety of different contextual situations captured by images. When a building is photographed alone the indexing process only takes account of the elements that value it in the composition of the image. But the technical choices made by the author of the image, such as the framing, and the light, must also be represented in the indexing.

Approaching this issue from a knowledge organization perspective enables a closer look at some contemporary issues specifically regarding the indexing of images (Dell Valle Gastaminza 1999; Moreiro Gonzalez and Robledano Arillo 2003; Smit 1996; Robledano Arillo and Moreiro Gonzalez 2002; Baca 2002). It is important to consider that the process of indexing iconographic documents as photographs and diapositives includes two levels of translation, firstly the translation from image to word and secondly from the word chosen by the indexer or the tag chosen by the user, to the term of the controlled vocabulary.

According to Layne (2002), an image cannot be described generically, and it is necessary to specify it, because it is always an image "of something". At the same time, the indexing process must be considered according to three different aspects: denotation – what appears in the picture, which will be recognized unequivocally by both the indexer and the user; connotation – what the image suggests, which refers to the emotional content of the message, giving meaning to its cultural aspects; and context – the conditions in which the image was produced and its narrative, involving issues such as how, when, where, why and who.

The image of the São Paulo Metropolitan Cathedral or Sé Cathedral in Figure 1 below is an architectural image, specifically a religious building, so we could use terms in the indexing process such as "architecture", "religious buildings", "churches" "cathedral". But Layne (2002) recognizes that "describing or identifying a particular subject will not necessarily serve all researchers of an image." Research circumstances may lead one user to seek "religious constructions" and another to search for "churches" while the circumstances of indexation may lead to the assignment of subjects such as "palm trees", "squares", "Se square".
Figure 1: São Paulo Metropolitan Cathedral or Sé Cathedral (Sakata 1997)

Unlike the text that may refer to "religious buildings" in general, the image will always be of a certain religious building and the indexing of the image must be done within a range of terms that allows for the retrieval of information. This range of terms referred to by Layne (2002) can be understood as the semantic field that will bring together terms with similar characteristics which allow their inclusion in a certain category and establishes an associative relationship between them. The attribution of a specific term to the image allows the inference of its generic term, when the indexer recognizes the similar characteristics of the term chosen with others included in the same category. At the same time, the many different characteristics establish the boundary of the semantic field and of the information retrieval. Therefore, the indexing process for architectural images in ARQUIGRAFIA must consider the connotation, denotation and context of each image, and their indexing terms must be understood as part of the controlled vocabulary whether they come from the library's subject lists or from the list of tags assigned by users.

4. Metadata and semantic interoperability

Metadata is essential in the knowledge organization process to search, access and retrieve information on the web because it reduces ambiguities and enables greater consistency and quality in knowledge organization systems. In ARQUIGRAFIA, the metadata was identified considering descriptive, technical and content aspects, as well as aspects related to its representation as cultural heritage on the web.

In the cataloguing process for architectural images it is necessary to respect their specific characteristics and contexts as representations of architectural works and
cultural objects that can be understood as artefacts of open-air museums. Thus, the procedures for the representation and retrieval of this digital cultural heritage required the analysis of existing cataloging standards to guarantee the access and sharing of information between institutions or individual users in this web collaborative environment.

The cataloguing standards such as AACR2, Anglo American Cataloging Rules (2004); ISBD (NBM), International Standard Bibliographic Description: No Book Materials (International Federation of Library Associations and Institutions 1987); VRA Code (Library of Congress and Visual Resources Association 2007); CDWA, Categories for the Description of Works of Art (Baca and Harping 2016) and CCO, Cataloging Cultural Objects (Baca et al. 2006) provide definitions and a formal structure to describe concepts, implicit and explicit relationships used in cultural heritage documentation, including formats for content representation from controlled vocabularies such as the Simple Knowledge Organization System and OWL (Ontology Web Language).

VRA Core and CCO are the tools that have presented the best results for describing architectural images in the descriptive data and content matching tests performed in ARQUIGRAFIA, because they allow an image to be catalogued as a work of architecture as well as registering its visual substitutes, whether digital or not.

Therefore, the team of researchers have been searching for a solution to consolidate standards which meets both the international requirements for syntactic and semantic interoperability, and the local needs for organization and access to information.

For the cataloguing of architectural images, data modeling identified the following fields: title, classification number, country, state, city, address, project author, year of construction, author of the image, date of the image, collection, album, register number, support, license Creative Commons; thematic metadata such as subject, image description and tags, as well as technical metadata such as width, height, resolution, bit depth and color representation of images.

As regards semantic interoperability it seems to be more viable when based on the construction of the intensional definition for each term of the controlled vocabulary in order to ensure the setting of the limits of the semantic field so that these definitions can be included as metadata in the modelling of interoperable ontologies that enable knowledge organization in the domain as well as information retrieval.

5. Conclusion

Our main goal is to describe the specific nature of the ARQUIGRAFIA collaborative web environment in order to preserve and disseminate a digital collection of Brazilian architecture which enables the convergence of institutional (public and private) and personal collections within a single free, open access website. Moreover, it provides a
singular opportunity to take an accurate look at contemporary KO issues, specifically regarding the construction of controlled vocabulary, image indexing, interoperability among systems and users’ active participation.

Regarding the construction of controlled vocabularies, we present an application model of terminological standards with the construction of a controlled vocabulary standard based on the theoretical reference of Dahlberg's concept theory, focusing on the definitions of each of the terms of the architecture domain to delimit the semantic field and the relationships between the terms, whether they come from the list of subjects from the library or tags assigned by users.

The indexing process for the architectural images in ARQUIGRAFIA must consider that an image cannot be described generically, and it is necessary to specify it, because it is always an image "of something". At the same time, the indexing process must be considered according to three different aspects: denotation; connotation and context, whereby its narrative must involve questions such as how, when, where, why and who.

Procedures for the cataloguing and retrieval of digital heritage in collaborative web environments were made in order to ensure access to and sharing of information between all users, whether institutional or individual. The required fields for data administration in the virtual environment were integrated and the set of metadata established.

User participation results in a set of tags that can contribute greatly to the creation and management of this digital collection, as it is carried out in a collaborative way, distributing resources, activities, and reducing costs.

From a conceptual point of view, a close look at some of the contemporary issues facing knowledge organization enables us to offer support and free help for other domains of knowledge to build their own collaborative visual environments on the web. The technological challenges concerning the design and operation of this collaborative web environment allow us to characterize ARQUIGRAFIA as an online experimental laboratory and a case study on the opportunities and risks of digital humanities projects.

References


Knowledge Organization Systems and cultural interoperability in open humanitarian settings

Abstract
This paper aims to identify the landscape of knowledge organization and representation applications developed in the open humanitarian data movement and consider how the high level of cultural interoperability can improve information exchange, the evaluation of humanitarian needs and the prioritization of humanitarian responses. Firstly, the paper presents reflections on the production, manipulation, and dissemination of open humanitarian data. Secondly, it discusses how knowledge organization meets the growing need to organize emerging forms of digital information. The case of Rohingya, a humanitarian disaster of historic proportion, is used to illustrate multilingual, cross-cultural, and social issues in knowledge organization and representation in humanitarian settings.

1. Introduction

Knowledge organization systems (KOS) have long been used as a means of information exchange and have functioned as standards in knowledge organization and resource discovery. The diversity of digital practices and the importance of knowledge consistency and reuse within systems in the cross-cultural and multicultural environment has attracted the attention of many authors and is becoming crucial for KOS and their interoperability (Barát 2008; García Gutiérrez 2002; Hudon 1999; López-Huertas 2008, 2016; Li Liew 2004). The need for global and local access to information in any language in any culture is an essential aspect of ethics in KO. Some authors in the literature have called for culturally warranted intercultural ethics and the goal of respecting cultural and linguistic diversity (Adler et al. 2016; Beghtol 2002, 2005; Hudon 1997; Mustafa El Hadi 2015; Tran 2017).

In the humanitarian domain, KOS enable humanitarian organizations to address needs with a better-coordinated response by improving information exchange, the representation of information needs, data relevance and overall effectiveness of retrieval. Nevertheless, attempts to standardize humanitarian data, and build a crisis ontology system have shown limited success (Di Maio 2007; Jihan and Segev 2013). It is common for agencies or information management officers to have to manually compile humanitarian data to fit specific requests (Warner and Obrecht 2016). Kessler and Hendrix noted that any proposed solution to improve coordination across agencies “should not require the replacement of existing information management systems, but rather focus on interoperability between existing systems” (Kessler and Hendrix 2015, 10). This paper firstly gives a brief overview of the recent open humanitarian data movement. It then discusses the use of KOS to support humanitarian data, the importance of interoperability, and the need for equal treatment of different languages,
communities, and cultures. The case of Rohingya refugee crisis is used to illustrate some multilingual, cross-cultural, and social issues in knowledge organization and representation in humanitarian settings.

2. Open humanitarian data and knowledge creation

In the humanitarian community, a significant amount of data remains difficult to store and process due to inconsistent schemes, ambiguous concept representations, and unstructured information. Open data, resulting from the global interconnection which is transforming Internet users into data providers, is nonetheless changing the face of humanitarian response and action. What drives the emerging phenomenon of digital humanitarianism is the desire to have un-annotated data better prepared and to create meaningful relationships among information resources (Meier 2015). Its motive is to allow machines to process, infer or combine information sources automatically, and transform them “into useful knowledge representation that depicts the actual crisis needs” (Jihan and Segev 2013, 526).

Recent projects in the open humanitarian data movement such as OpenStreetMap\(^1\) show how civil society groups are pushing for open platforms and frictionless data mobilizations. Where digital data remains inaccessible, citizens organize themselves to reconstitute information, which is the case of Open Knowledge International\(^2\). Humanitarian Data Exchange (HDX)\(^3\), an international open data repository for crisis and humanitarian needs, is an emblematic case of offering homogeneous documentation whose structure has been preserved. This platform employs Humanitarian eXchange Language (HXL), a data standard designed to help improve the sharing and consolidation of data across agencies and move the humanitarian community past its reliance on Excel spreadsheets. However, the still early and uneven nature of the humanitarian data-sharing process in different countries show that the problems of ambiguity, inconsistency, and incompleteness have not been solved. Indeed, knowing to identify, reuse and exchange data, and transform it into knowledge representation that describes the current crisis needs to become a condition for the efficiency of decision making in situations of emergency, natural disaster, and conflict.

3. The use of KOS to support humanitarian data

The increasing growth of information, information media, channels, and services leads information professionals to cope with the new needs by developing innovative methods for indexing the ever-growing digital contents and adapting them to new information carriers and repositories. The integration of KOS in digital humanitarian environments would facilitate the integration of the established corpora of recorded

---

1 See https://www.openstreetmap.org.
2 See https://okfn.org/.
3 See https://data.humdata.org/.
knowledge in humanitarian action and digitally generated resources in the emerging open data ecosystem. Publishing of KOS on the Web opens many opportunities for exploiting KOS that were not straightforwardly available:

- KOS provide a standardized way of expressing data and make meaningful relationships between data points. They help improve information flow for decision-makers, helping them allocate resources for response activities.
- KOS record knowledge orders and structures. They can thus relate obsolete knowledge and deprecated terminology with current activities, enhancing evidence and learning each time new knowledge is generated.
- KOS offer terminology hints, evidence of cultural and sociological phenomena from the past. In this way, KOS can raise awareness of potential users and their needs and improve humanitarian performance.

Universal knowledge classifications, as “language independent concept schemes” (Slavic 2011, 23), can assist humanitarian responders in presenting information and formulating queries. The idea behind SKOS and standards such as ISO/IEC 13250 Topic Maps and BS 8723 Structured Vocabularies for Information Retrieval is that “the availability and exchange of controlled vocabularies of various kinds, and subject indexing languages in an open networked environment may contribute to resource discovery through referencing, resolving language ambiguities and providing semantic context for text processing” (Slavic 2011, 44). The application of universal knowledge classification could contribute to not only better resource discovery in emergency response and relief coordination but also improving terminological and research tools for knowledge transfer between clusters/sectors such as Education, Health, Shelter, Water and Sanitation, among others. In the next section, the Rohingya humanitarian case is used to illustrate certain important points in the representation of knowledge and resource description terms in the context of cultural and linguistic diversity.

4. Integration of knowledge: the Rohingya case

Since violence broke out against the Rohingya in 2012, more than half a million Rohingya Muslims have fled their homes in Myanmar’s Rakhine province. In October 2017, the UN Refugee Agency (UNHCR) described the Rohingya’s situation as the “world’s fastest growing refugee crisis.”4 The Rohingya case can provide a context to highlight intervening factors that may influence ethical decisions made by the information professional, namely cultural patterns, social usefulness, social responsibility, respect for oneself, respect for other individuals and institutions, among others (Froehlich 1994). On the HDX platform, there are 22 datasets related to Rohingya displacement; 11 of which are geographical data. In data.world5, there are

5 See https://data.world/.
seven projects and datasets related to health facilities in refugee sites, need and population monitoring, and settlement infrastructure. Below, we examine some issues concerning resource description and discovery that classification schemes ought to take into account.

Stateless status

Under Myanmar’s 1982 citizenship law, a million Rohingyas were stripped of citizenship, leaving them *stateless* and without access to education, healthcare, freedom of movement, and liberty, among others. In Myanmar, the term “Rohingya” itself is taboo and Buddhist leaders refer to Rohingyas as “Bengali” instead. The term “stateless,” as well as “resident foreigners” and “associate citizens,” should be distinguished while constructing or consulting classification schemes, as the legal differences between them can have life-changing consequences.

Refugee, migrant, internally displaced people (IDP)

According to the UN High Commissioner for Refugees, hundreds of thousands of Rohingyas have sought refuge in nearby countries; the vast majority of them migrate to Bangladesh. The question of classifying someone as a refugee, not just a migrant, has complex legal implication and is extremely important because refugees have a right to be protected under international law and cannot be forced to return home.

Another marginalization issue in this displacement is the category of IDP, which has received scant international attention. According to UNHCR, the UN Refugee Agency, “IDPs stay within their own country and remain under the protection of its government, even if that government is the reason for their displacement”\(^6\); therefore, they are among “the most vulnerable in the world.” Some widely used classification systems have decided not to include this concept. For example, the Library of Congress Subject Headings (LCSH) does not contain “internally displaced people” but only “internal migrant” (Library of Congress 2017), which can only be seen as a term related to the former.

---

Figure 1: The heading “internal migrant” in the 39th edition of LCSH

<table>
<thead>
<tr>
<th>Internal migrants (May Subd Geog)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Here are entered works on persons who change their place of residence from one locality to another within the borders of a single country. Works on foreign-born persons who enter a country intending to become permanent residents or citizens are entered under Immigrants.</td>
</tr>
<tr>
<td>UF In-migrants</td>
</tr>
<tr>
<td>Migrants, Internal</td>
</tr>
<tr>
<td>Out-migrants</td>
</tr>
<tr>
<td>BT Persons</td>
</tr>
<tr>
<td>RT Migration, Internal</td>
</tr>
<tr>
<td>NT Children of internal migrants</td>
</tr>
<tr>
<td>Church work with internal migrants</td>
</tr>
<tr>
<td>Women internal migrants</td>
</tr>
</tbody>
</table>

Regarding activities of knowledge organization and representation, some authors have raised concerns about the bias in the representation of concepts related to prejudice and discrimination (Berman 1993), the negligence of the indexer in choosing terms (Dahlberg 1992), the non-neutrality of the subject and undeniable damage to the user (Milani and Guimarães 2011). A term such as “internal migrants” can carry different connotations in humanitarian action than in urban education or labor relations. A synthetic approach, “placing domain analysis within the context of a comprehensive classification” incorporating interdisciplinarity, can be considered in coping with the challenge of semantic ambiguity (Szostak, Gnoli, and López-Huertas 2016, 145).

Islamic practices and social customs of the Rohingya

The Rohingya’s religious orientation (Muslim) and racial antecedents (Bengali) had not been recognized by Myanmar’s leaders whose majority are Buddhist and speak their own dialect (Agrawal 2018). Extreme levels of discrimination coming from extremist groups is a break from how Rohingya communities used to be treated. To avoid prejudice and errors of bias, we should take into account the socio-cultural context. Multiple authority identifiers and cross-references can be used to reflect the languages used in addressing these topics more accurately. The preliminary assessment of socio-normative values, beliefs, and practices of the Rohingya (see Table 1) is vital to recommending appropriate humanitarian responses regarding health, wellbeing, and the protection of children, adolescents and women (Ripoll 2017).
Table 1: Examples of terms and variations related to Rohingya Islam

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Variations</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Health and well-being</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>jinn</td>
<td>djinn (French), dschinn (German), dzinn (Polish); genie (related term)</td>
<td>Spirits in Islamic mythology</td>
</tr>
<tr>
<td>wudu</td>
<td>özu (Bengali); Lamaz ecar (Chechen); abdest (Albanian, Polish, Turkish); woedoe (Dutch); ablution (related term)</td>
<td>Obligatory ritual washing</td>
</tr>
<tr>
<td>kabiraj</td>
<td>boiddah, hakime (Rakhine dialects)</td>
<td>Local healer who uses herbal medicine; doctors who heal jinn possession (Ripoll 2017)</td>
</tr>
<tr>
<td><strong>Protection of children, adolescents and women</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>purdah</td>
<td>pardah (variation); parda (German); paranja (related term)</td>
<td>Practice of female seclusion</td>
</tr>
<tr>
<td>burqa</td>
<td>burkha, bourkha, burka, burqua, burqu’ (variations); boerka (Dutch)</td>
<td>Veil to conceal the face and body</td>
</tr>
<tr>
<td>hijab</td>
<td>hidschab (German); hidzsáb (Hungarian); tesettür (Turkish); headscarve (related term)</td>
<td>Veil to cover the head and neck but leaves the face clear</td>
</tr>
<tr>
<td>waqf</td>
<td>vakuf (Bosnian, Slovenian); wakf (Polish); vakf (Czech)</td>
<td>Charitable endowment under Islamic law</td>
</tr>
<tr>
<td><strong>Social customs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>jumu’ah</td>
<td>jum’aa (Bengali); yumu’ah (Spanish); cuma namazi (Turkish)</td>
<td>Friday prayer</td>
</tr>
<tr>
<td><strong>Social solidarity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>zakat</td>
<td>zekat (Bosnian); azaque (Spanish); zacate (Walloon)</td>
<td>Form of mutual aid to help the poor</td>
</tr>
</tbody>
</table>

These above examples show the importance of language translation in disaster response and the role of KOS in producing an improved mode of communicating with communities and people affected. From a cultural perspective, knowledge organization processes, tools, and products “suffer a direct impact from cultural diversity” (Guimarães 2017, 92). The tension between Otlet’s universalist vision and the culturally anchored “local” remains a fundamental theme in classification, authority control, and links for expanding resource discovery from a local perspective to the global environment. Use of multilingual authorities and multiple authority identifiers
can be helpful as a means of structuring information in the intercultural interface. KOS tool design and implementation implies considering the integration of multiple authorities, through multilingual authorities and multiple authority identifiers, to accommodate the cultural warrants necessary to assure equal treatment among communities and cultures.

5. Conclusion

The pervasive power of digitization and open data causes scientific, educational, economic and cultural communities to change modes of accessing, sharing and disseminating knowledge. In the humanitarian domain, the goal of open data initiatives is to improve information exchange during extreme situations and the effectiveness of humanitarian response. Although significant efforts have been made regarding the systemic and semantic heterogeneity of data, culture heterogeneity in this domain remains largely unaddressed. We need to take into consideration cultural diversity in designing KOS and modes of access to humanitarian knowledge.

The Rohingya case is presented to illustrate certain multilingual, cross-cultural, and social issues in the representation of knowledge in the context of cultural and linguistic diversity. It was chosen not only because of its humongous scale and multidimensionality in terms of humanitarian needs, but also the implications of the cultural confrontation that took place.

This paper emphasizes the importance of an information architecture which could serve as a model in local KOS, and their interoperability with global KOS. It will be a step towards the increasing participation and interaction of different knowledge communities in improving KOS terminology and enhancing integration of knowledge in the humanitarian sector.

References


Dorota Siwecka

Knowledge Organization systems used in European national libraries towards interoperability of the semantic Web

Abstract
The paper presents how European national libraries use knowledge organization systems in projects connected with providing open access to their bibliographic and authority data on the Internet. Over the last ten years there have been two main trends related to the implementation of these projects in order to bring well organized bibliographic and authority data to the world of the Semantic Web. The first trend has been to publish national bibliographic and authority data as open data freely available to everyone on the Internet. The second new approach has been to link open data with other data available on the Web. Performing these kinds of tasks of opening and linking bibliographic and authority data, libraries are involved in creating new tools, among them ontologies, schemas and dictionaries, as well as improving those already existing.

There is no doubt that technological development has contributed to the revolution of information retrieval methods, especially in the area of managing connections between data. Some changes have impacted on electronic catalogues, allowing users to search for information via a larger number of access points than simply author name or title (as in alphabetical catalogues) or subject heading (as in subject catalogues). Depending on the capabilities of the system, we could search for information about a collection using such search terms as date, place or publication language. Electronic catalogues are able to quickly create a list of all the works by a given author in chronological order. However, it is still not possible to link information regarding differing ways of presenting the same title, e.g. in the form of text, film or theatrical adaptation. Further retrieval mechanisms in both traditional and OPAC catalogues have depended to a large extent on references or users being expected to initiate another search query.

The FRBR (Functional Requirements for Bibliographic Records) idea has, obviously, been known within the librarian community for at least 20 years. So far, however, the technical limitations of library systems have not allowed the indication of relations between the various elements of a collection. Linked open data technology and the RDF (Resource Description and Framework) standard model have created a real opportunity to make changes in this area.

New standards implemented for global Internet could be the perfect solution for problems facing libraries. Over the last decade the development of libraries has been, to a large extent, determined by ontologies, thesauri, classifications and other systems of knowledge organization. These tools organize the terminology used in online library catalogues, bibliographic databases and digital libraries, as well as relations between terms. For now, however, only a few solutions are providing high quality data to enrich
the Semantic Web, which is likely to become one of the most important technological solutions in the near future.

Libraries have been involved in the development of various knowledge organization systems (KOS) for decades. These actions have led to the creation of authority files (subject, names or places), as well as information retrieval languages (e.g. subject headings, classifications, thesauri) or collections of structured bibliographic data presented in the form of standardized descriptions (ISBD). According to international guidelines, national libraries should create an inventory of best practices for other libraries to follow. That is why they are probably the most appropriate institutions to transfer best library solutions to the Web – the main source of information in the present day.

How do national libraries in Europe support the development of the semantic Web?
According to Tim Berners-Lee, the inventor of the World Wide Web and Linked Data initiator, there is a 5-star deployment scheme for Open Data. The highest level is linked open data (Hausenblas 2012). This kind of information will probably lead us to the Semantic Web (where machines can “understand” information content). One of the key tasks of national libraries is to create well organized information - in the digital age they should take advantage of being visible on the Internet offering high quality information by connecting information from their resources to existing popular ontologies. Through their participation in projects aimed at opening and linking bibliographic and authority data, national libraries are involved in creating new tools, among them ontologies, schemas and dictionaries, as well as improving those already existing.

Case studies of individual projects carried out by national libraries in Europe1 indicate two general categories: 1) projects aimed at publishing data in open access (Open Data) for re-use (see Table 1) and 2) projects focused on publishing different kinds of data as Linked Open Data (see Table 2) – there are 12 national libraries which have decided to share their data this way: the British Library, the National Libraries of Finland, France, Germany, Greece, Hungary, Ireland, Latvia, Netherlands, Portugal, Spain, Sweden, Wales and the National Central Library in Florence.

---

1 In the survey, national libraries from 50 European countries were examined. Information about all the projects undertaken were collected, mainly from the libraries’ official websites and other materials available on the Internet (blog posts, slide shares, articles etc.). Other sources were: Open Knowledge International, Datahub. The easy way to get, use and share data [online]: https://old.datahub.io/ [access: 31.01.2018], European Data Portal [online]: https://www.europedataportal.eu/ [access: 31.01.2018] and The Linking Open Data cloud diagram [online]: http://lod-cloud.net/.
Table 1: Selected examples of Open Data projects implemented by national libraries in Europe (as at 31-12-2017). The author’s own research

<table>
<thead>
<tr>
<th>National library</th>
<th>Project</th>
</tr>
</thead>
</table>
| 2. National Library the Netherlands | • Early Dutch Books Online (EDBO or DPO) (Dutch only) in cooperation with the university libraries of Amsterdam (UvA) and Leiden  
• Medieval Illuminated Manuscripts in cooperation with Museum Meermanno | Huis van het Boek  
• The Dutch Digital Parliamentary Papers (Dutch only) in cooperation with the House of Representatives  
• Watermarks in Incunabula printed in the Low Countries |
| 3. National Library of Poland | data.bn.org.pl |
| 4. The British Library | • British National Bibliography  
• British Library Integrated Catalogue – Books  
• British Library Printed Music  
• Alice's Day |
| 5. Swiss National Library | • Handschriften Rainer Maria Rilke  
• The Gugelmann Collection (Schweizer Kleinmeister)  
• Swiss Book (2001 - )  
• Max van Berchem image collection  
• Eduard Spelterini image collection |

Table 2: Selected examples of Linked Open Data projects implemented by national libraries in Europe (as at 31/12/2017). The author’s own research

<table>
<thead>
<tr>
<th>National library</th>
<th>Project</th>
</tr>
</thead>
</table>
• RAMEAU subject headings (STITCH) |
GND Ontology – https://d-nb.info/standards/elementset/gnd |
| 11. National Library of Sweden | Libris |
Both in the first and second general categories, national libraries have decided to publish data using free licenses. Various types of high quality data sets have been opened; bibliographic and authoritative files (of subject, geographical and personal names). These data in some way organize the knowledge available on the Internet and can thus be considered as examples of KOS.

Two subtypes can be distinguished in the process of providing open data: a) publishing open structured data in formats typical for libraries (e.g. MARC, MARCXML, Dublin Core) and making them available using the SRU, OAI-PMH or Z39.50 protocols, and b) publishing open data stored in a format conducive to the development of the Semantic Web. The first subtype of data is actually offered by the majority of national libraries in Europe. In addition to the typical bibliographic data from library catalogues or national bibliographies, there are also projects that provide data from specific library collections (e.g. Early Dutch Books Online collections, Medieval Illuminated Manuscripts, The Dutch Digital Parliamentary Papers of the National Library of Netherlands\(^2\), Collection on the I World War, Collection of digitized newspapers of the National Library of Portugal\(^3\), Scottish Bibliographies Online, the Soviet Posters of the National Library of Scotland\(^4\)). There are also files made available in an open format, e.g. CSV (an example of this can be records from the British Library: Alice's Day, Black history Month\(^5\) and collections of the Swiss National Library\(^6\) Handschriften Rainer Maria Rilke, Bildersammlung Annemarie Schwarzenbach). The second subtype distinguished in the process of providing open data is related with publishing open data stored in a format conducive to the development of the Semantic Web (e.g. British National Bibliography provided in RDF/XML format). Bibliographic open data on the Internet have also been made available by Swiss National Library (Swiss Book). Via the portal data.bn.org.pl, the Polish National Library gives access to bibliographic and authority records in JSON, XML, MARXML and MARC formats. 

Apart from providing bibliographic and authoritative data, national libraries in Europe also make their own ontologies available online as open data. Analysis of the examples shows that the greatest achievement in this field has been the Finnish National Library, which provides access to nearly forty ontologies, classifications and thesauri

---

\(^4\) These and many more datasets are available on the National Library of Scotland website: https://data.nls.uk/. Accessed 31.01.2018.
on the special service Finto. Published on its portal, knowledge organization systems are created independently by the Finnish NL (e.g. Suomalaiset yhteisönimet – Finnish Corporate Names), in cooperation with other organizations (e.g. JUHO – Julkishallinnon ontology – Finnish Ontology for Public Administration created jointly with the Office of the Prime Minister, Semantic Computing Research Group (SeCo) and the State Treasury), or by external institutions – Finnish NL in such cases being only a publisher of datasets (e.g. Iconclass, Julkisten palvelujen luokitus – Classification of public services, Korkeakoulujen tutkimustiedonkeruussa käytettävän tieteenalaluokitus – scientific discipline classification). Finto, however, is not only a database of ontologies and thesauri (mainly Finnish, but also including international collections) which can be used on the Internet. The service also allows users to browse and search all ontologies, as well as download datasets, usually in the SKOS/RDF formats: RDF/XML and TURTLE. It also provides an API enabling the use of these thesauri and ontologies in other applications (Kansalliskirjasto n.d.).

As mentioned above, another example of involving national libraries in the development of the Semantic Web is the creation of bibliographic and authority data, dictionaries, thesauri and ontology as linked open data (LOD). When analysing LOD projects, more attention should be paid to two aspects of using knowledge organization systems. First, KOSs are used in the process of developing the data description model. Second, national libraries are using KOS functioning on the Semantic Web to link their data to them.

Considering the first aspect, the most common solution is to apply existing standards wherever possible. According to an OCLC survey conducted in 2014 and 2015 among 112 linked data projects and services (including 14 national libraries around the world), the most frequently used ontologies are the Simple Knowledge Organization System (SKOS), Friend of a Friend (FOAF), DCMI Metadata Terms (DCterms), Dublin Core Metadata Element Set (DCE), Schema.org vocabulary (schema.org), The Bibliographic Ontology (BIBO), Local Vocabulary, VOCABS rda, Europeana Data Model vocabulary (EDM), ISBD elements (ISBD), WGS84 Geo Positioning (GEO), BIBFRAME Vocabulary (BF) (Smith-Yoshimura 2016). The most frequent standards in the national library projects analysed were: SKOS, BIBO, DublinCore, OWL (Web Ontology Language) and FOAF (see Chart 1). In addition, when developing data models, some entities decided to develop their own schemes, due to insufficient representation of relations or terms available in the above-mentioned schemes, e.g. British Library Terms RDF Schema (British Library 2017).

Among the projects analysed, there were both simple models, using a small number of (up to 4) ontologies, and those that required the development of their own more

---

complex data models (e.g. the British Library Data Model developed for books, magazines, and publishing announcements).

![Chart 1: Number of LOD projects from European national libraries that have been using specific patterns for data modeling in 2008-2017. The author’s own research](chart)

The second aspect regarding exploitation of KOS in European national library projects considers creating links to other KOS existing on the Web. The main reason for publishing and linking data is to expose them to a wider audience on the Web (Smith-Yoshimura 2016). According to the OCLC survey, the data are most commonly associated with such projects as (in order of popularity): VIAF, DBpedia, GeoNames, and id.loc.gov (Smith-Yoshimura 2016). The case study of European national libraries shows that in the library environment the above projects are also the most popular (see Chart 2). Links to other library collections such as main catalogues, union catalogues (e.g. Sudoc, Libris and OCLC WorldCat) and digital libraries are also very common. The Spanish project datos.bne.es links, for example, to Wikipedia, while the Swedish union catalogue LIBRIS links to the Biblioteca Digital Hispánica and to the French portal data.bnf.fr. Other projects link to Sudoc, WorldCat, Catalog Collectif de France, Gallica, the catalog of the French National Library, the catalog of manuscripts, the archive of BNF or Dewey.info.

Linked open bibliographic data are available on the Web at the German National Library (Titeldaten der Deutschen Nationalbibliothek) and the British Library (British National Bibliography). Authoritative data in the form of linked data projects have been
released, among others, by the Greek National Library (National Library of Greece Authority Records\(^8\)) and the German National Library (Gemeinsamen Normdatei).

Apart from linking and opening bibliographic data, there are also projects which focus on linking ontologies, vocabularies and other types of KOSs. An example is Linked Logainm, which contains data related to Irish geographical names that enrich Irish library metadata. This is a unique collection of information on Irish-language geographical names. The National Library of Ireland decided to develop its own dictionary due to the need to specify some of the terms. Linked Logainm therefore contains more detailed information about Irish geographical names than DBpedia or Geonames (Lopes et al. 2013). The data in the Linked Logainm project are linked with such sets as Europeana, DBpedia and Geonames.

Additionally, the finto.fi service, described above, provides not only open ontologies and thesauri, but also links from them to other projects (for example, YSO links terms with the KOKO ontology, YSA, Allärs, LCSH; Metatietosanasto links to rdaregistry.inf and ISBD Media Type).

Chart 2: Number of LOD projects from European national libraries linking to individual resources on the Internet. The author's own research

![Chart 2](image_url)

Another example of organizing network content is the creation of ontology

---

collections. In addition to the already mentioned Finto service, the National Library of Finland combined all the most important Finnish ontologies within one KOKO collection. The ontologies include the Finnish General Upper Ontology YSO and ontologies that extend and refine YSO such as the Ontology for the Museum Domain, the Ontology of Applied Arts, and the Finnish Ontology of Photography (Kansalliskirjasto 2018).

Undoubtedly, the idea of open linked data is becoming more and more popular in the library environment. Among national libraries in Europe, the number of LOD projects increased from 1 in 2008 to 16 in 2017. Almost every year there are from 0 to 3 new projects implemented by those libraries (see Chart 3).

Chart 3: Number of LOD projects implemented by national libraries in Europe in 2008-2017. Authors own research

It is noteworthy that starting and maintaining the standalone projects is not the only way of redistributing library data on the Web. Many national libraries (both those that started their own projects and those that did not) assigned a Europeana Data Exchange Agreement to provide their data under open metadata licence CC0 to Europeana’s Linked Open Data project: Austria, Denmark, Serbia, Portugal, Czech Republic, Luxemburg, Slovenia, Ireland, United Kingdom (‘Linked Open Data Means Sharing Knowledge’ 2012; Haslhofer and Isaac 2016).

Another example of cooperation is supporting the German National Bibliography with data from the Swiss National Library (Schweizerischen Nationalbibliothek NB

9 The National Library of Serbia enriched the Europeana collection with records from its doiSerbia service (Gardašević 2013) and has also launched its digital library Veliki rat (Great War) with "Europeana Collections 1914-1918" project (Europeana 2017).
There are also some projects focusing on releasing data freely from national libraries to Wikidata and Wikimedia, e.g. National Library of Wales connected their *Welsh Landscape Collection* with Wikidata. National Library of Netherland put all its freely available images in categories: Koninklijke Bibliotheek on Wikimedia Commons (Koninklijke Bibliotheek n.d.).

**Conclusions**

The popularity of LOD projects in the library environment will probably increase. Projects from other types of libraries (e.g. scientific, public) should be examined in order to get the full picture of the development of libraries and linked open data (LOD) in Europe. The growing number of publications devoted to the issue of linked open data for culture is also evidence of the constantly growing interest in this subject.

Certainly, the biggest drawback during the research into projects implementing the idea of the Semantic Web by national libraries in Europe was the lack of clearly defined standards for describing such projects. In fact, each project is described according to different criteria. Searching for information should be expanded to various post-conference presentations, blog posts, articles and other resources. It was often a problem to determine whether a given project, as referred to in the materials found on the Web, is still being continued by the institution.

Perhaps the best solution would be an international standard for the description of LOD projects. The best described project is the *British National Bibliography as Linked Open Data* portal, which besides technical documentation also provides detailed instructions for searching data and FAQ (British Library n.d.). The only thing missing is the date when information was published and last updated.

The international standard for describing LOD projects undertaken by libraries (not only national) could be the basis for an international database updated annually and collecting information from libraries. There are of course many different websites which include links e.g. to the BNF, BL, BNE and DNB projects – but there still remains a lack of up-to-date information. The standard description could contain basic information such as: project title, schemas used, vocabularies, license, start date, data type (bibliographic, authority, etc.), data source (for example, library catalogue, national bibliography, digital library collection, etc.), information about how to access and download the data and contact information with link to a website providing more information about the project.

Chart 3 shows that the number of projects is constantly growing, thus such a database would be a perfect source of information for other libraries which are interested in releasing linked data. It could also serve as a teaching aid for library and information science teachers and students.

If we want to take part in creating the Semantic Web and also be visible on the
Internet for users who are searching our data sets to use information in their applications or to connect their data with ours, then besides linking their open data, libraries should also take care of describing their data for the benefit of real users.

References


Andreas Oskar Kempf

The Need to interoperate: structural comparison of and methodological guidance on mapping discipline-specific subject authority data to wikidata

Abstract
The linking paradigm of Linked Data (LD), the RDF-based information architecture of the WWW, and RDA with its underlying Entity-Relationship-Model (ERM) demand an increased entity-based semantic interoperability of subject authority data. Wikidata (WD), the knowledge base and sister project of Wikipedia (WP) seems to be a promising environment for joint efforts to bring authority data into the semantic web and to undertake this task. However, it seems necessary to bear in mind that, with regard to subject authority data, WD interlinks different forms of knowledge organization (KO) that are characterized not only by different functional purposes but also by distinct structural principles of modeling concepts clearly linked to the diverse information systems (IS) from which they derive. On these grounds, this paper provides methodological guidance on how to approach a mapping process between topical thesaurus concepts and WD items.

1. Introduction
As the web has expanded in syntax and scope, it has evolved from what was originally a web of documents into an all-encompassing medium (see Gradmann 2013) and has become the place to search for information. At the same time, the so-called ‘Web of Things’ enables the representation of everything that could become part of an RDF statement. This has enormously extended the scope of representation in the web and “finding and selecting moves steadily toward connecting and relating every record with every other record in an all-embracing web” (Buckland 2017, 176).

In the spirit of the LD paradigm new web environments have evolved to interlink different types of entities. While they seem suitable to interlink individual concepts, such as author names (Neubert 2017), when it comes to subject authority data, they run the risk of juxtaposing entities with different representation purposes and frames of meaning.

By referring to topical concepts of the STW Thesaurus for Economics as mapping candidates for WD items, this paper demonstrates how one might approach a mapping process between topical thesaurus concepts and WD. Section 2 states the research problem by hinting at the interconnectedness between functional and structural principles of KO entities and the respective IS to which they are tailored. Section 3 presents core functional requirements and structures of documentation languages (DL) in general, before referring to the concrete example of the STW. Section 4 refers to the function and structure of concept modeling practices in the WD backbone WP. Chapter 5 contrasts the concept modeling principles of the two forms of KO by referring to their corresponding IS purposes and parameters. On this basis, section 6 takes stock of the
ways in which one might select potential thesaurus concepts for a systematic mapping process between the two by referring to the latest ISO standard. The conclusion summarizes the role of WD as a linking infrastructure in this context.

2. Problem statement

The library field has long-standing expertise in the development of subject authority data. For many years, such vocabularies were applied within the confines of a single library collection. As the web has become the place to search for information, the scope of thesaurus-indexed collections is getting more and more relative. In recent years, controlled vocabularies have increasingly been published on the web to extend their coverage beyond mere bibliographic metadata, but their inherent structural characteristics, deriving from the IS they were initially developed for, have remained. However, LD and the RDF-based architecture of the semantic web demand an increased semantic interoperability of subject authority data to “get rid of” vocabularies that are “deeply rooted in the focus on information containers” (Gradmann 2013, 255).

Linked environments promise to bring together the same entities of different authority files, thus serving as a linking hub. However, while they seem suitable for individual authority data, a closer look at matching subject authority data in a crosswalk service like WD reveals clear differences in modeling concepts between the different forms of KO they relate to. By taking the example of the STW, we provide guidance on best practice for how to approach a mapping process between subject-specific authority data taken from a thesaurus and topical items from WD, paying special attention to their distinct functional and structural principles.

3. Controlled vocabularies for indexing and search

3.1 Structural characteristics and representation functions

According to the latest ISO standard, a thesaurus is a formally structured vocabulary in which “concepts are represented by terms, organized so that relationships between concepts are made explicit, and preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms” (ISO 2013, 14). Conceptual and terminological control to “guide both the indexer and the searcher to select the same preferred term or combination of preferred terms to represent a given subject” (ibid.) are thus two key and interlinked principles underlying thesaurus design.

This concept navigation function is always related to specific IS parameters. According to Wersig (1985), these parameters (further expanded in section 3.2 with reference to the STW) include the degree of stability of the domain-specific terminology and the terminological richness (i.e. the amount of descriptors and nonDescriptors). Additionally, he emphasizes different representation functions in the modeling process of a DL element – that is, a thesaurus concept represented by a
descriptor – and distinguishes four different levels. On a first level it could represent a class of concepts such as, for example, the STW descriptor “Oil and gas services”\(^1\). On a second level, the DL element could represent a particular class of objects from a set of object classes formed by the class of concepts. Which of these objects are specifically meant, in turn, does not derive from the verbal description of the class itself, but only from the context of the element. For example, the STW descriptor “Institutional infrastructure”\(^2\) could represent legal as well as culture-specific conditions of social and economic life. A third level represents the entire system-specific context, e.g. what kind of documents need to be indexed and according to which indexing rules. This could lead to a system-specific narrowing or broadening of the meaning of the DL element. On a fourth level, in practice the DL element represents the amount of concrete documents it was assigned to \((ibid., 74 \text{ p.})\).

Moreover, again referring to Wersig \((ibid. 77 \text{ p.})\), when conducting conceptual control, two aspects are of particular practical importance with regard to the semantic frame of a concept: first, the implicit and explicit meaning of a term, second, the natural language meaning and the DL meaning of a term. With regard to the first aspect: the implicit meaning(s) of a term is/are learned in the socialization process and come(s) out of its/their usage. (The) explicit meaning(s) of a term is/are contained in authoritative lexical works of an essentially encyclopedic nature. In the case of a DL they could be made explicit in scope notes. Both meanings do not need to be identical to the current use of a term for indexing. With regard to the second aspect, DL recruit or build their vocabulary from the vocabulary of currently used natural language, incorporating its communicative meaning \((i.e., \text{ the set of all uses of the term realized in current or past communication situations})\), its lexical meaning usually being a subset of it. Communicative/lexical meanings are not necessarily always fully accepted in the DL or taken over \((e.g. \text{ polysemy control})\). Equally conceivable is the reverse case, in which a meaning could be assigned to a term, even though it does not correspond to its lexical or communicative meaning \((e.g. \text{ quasi-synonym definitions})\)\(^3\).

In general, according to Wersig \((ibid. 79)\) conceptual control during thesaurus development and maintenance should take into account the following aspects: identification of the communicative meaning, whereby the lexical meaning cannot always be adopted into the DL; it often needs to be supplemented with meanings of current use in the relevant domain \((\text{conceptual analysis})\), elimination of \((\text{components of})\) meanings that are not relevant to the DL, by exclusions \((\text{conceptual adjustment})\), for example; extension of further \((\text{components of})\) meanings that are not communicatively defined, but are to be subsumed here for purposes of the DL and made

---


\(^3\) The fact that a thesaurus could be used in more than one IS \((ibid.)\) should be excluded here.
explicit (conceptual extension)\(^4\).

Integration of new concepts or their elimination could affect the definitions of meaning of other conceptual units, therefore monitoring of meaning as well as of assignment needs to be done. Moreover, the use of a term in the context of common language and discipline-specific terminology is not static (see ISO 2011, 96).

Structures for conceptual order in a thesaurus which give orientation and assist discovery are its classical semantic relations and an existing category system.

### 3.2 The STW thesaurus for Economics

The STW, with nearly 6,000 descriptors and about 20,000 non-descriptors, broadly covering most key subject areas including all those relating to economics, comprises different types of concepts. In its geographical subthesaurus, comparable to a name authority list, it mainly contains individual concepts, i.e. country descriptors, while the other six subthesauri include general concepts, characterized by different levels of abstractedness, culminating in the subthesaurus ‘general descriptors’\(^5\).

With regard to its IS context, merging four different DL, ranging from a simple keyword list to a fully-fledged thesaurus and reflecting specific collection focuses, marks the birth of the STW, which today is used in the IS of the respective institutions originally involved in its development, as well as in other IS. The large amount of source terminology has further promoted the terminological richness already present in this social science domain. This is reflected in extensive equivalence classes in some of its branches. Regarding key principles for concept modeling, both methods, precombination and post-coordination, are used. The overall thesaurus structure is polyhierarchical.

Right from the start, the scope of the STW, now available in German and in English has been constantly expanded. Similar to the situation in other disciplines, scientific discourses in many sub-disciplines of the economic sciences have become more and more international. The ZBW – Leibniz Information Centre for Economics, which has taken over STW development, has consequently directed its information services towards an international, English-speaking scientific community (Kempf/Neubert, 2016).

Comparable with other social science thesauri, the terminology is rather fuzzy and the degree of terminological stability is rather low. Some terms could stand for different theoretical models which cannot be further differentiated. Moreover, the concepts are often deeply embedded within scientific discussion processes. In addition, far-reaching discourse shifts and new subject fields can be regularly observed. In general,

\(^4\) Wersig (ibid., 79) also mentions system-related specifications when using a cross-system thesaurus.

developments in this domain concern concepts, terminology semantics, and vocabulary structure. For this reason, after more than 15 years of permanent, though rather isolated ad-hoc updating, the STW has been completely revised, descriptor by descriptor, over a period of several years. In the end, more than 750 thesaurus concepts have been added, and nearly 1,100 (out of approximately 6,000) concepts have been eliminated. In addition, entry terms have been adapted and the systematic structure has been improved (Gastmeyer et al., 2016). Detailed change reports, which make changes of the STW traceable, are provided using the published SKOS files of its latest versions (Neubert 2015).

Enhanced interoperability has already been achieved thanks to vocabulary mappings established in the past, allowing STW content descriptions to be translated into subject information using other vocabularies (and vice versa). An example is the continuous intellectual mapping between STW and subject authority data of the German Integrated Authority File (GND) (Kempf/Neubert 2016).

4. Wikidata & Wikipedia – information system parameters and concept structure

Launched by the Wikimedia foundation in the year 2012, WD serves as a shared knowledge base to provide structured data for the nearly 300 different language versions of WP. Hence, the collaboratively-edited free online encyclopedia WP is the starting point for WD, its concept modeling referring to an encyclopedic design. In the first place, its entities serve as reference in the search for definitions and factual information of encyclopedic nature. Furthermore, its structural form of KO is for learning; for user guidance, connections between entities are established through the use of categories.

Taking the open production model of WP and its fairly vague notability criteria into account, the level of specificity of an entry term can vary enormously. Which conceptual unit receives article status is up to the participating contributors. The conceptual scope of theme or topic units, that is, their entry lemmas, which decisively shape the structure of an encyclopedia, can be broad or narrow. Linked to this, the depth of explanation can also vary greatly, ranging from rough overviews to explanations which place their subject in a broader context. Related to this and depending on the content, the lifecycle and degree of dynamics of an entry also varies. Although its claim is universal and it contains all types of information, there are big gaps in the list of lemmas which are edited by the Wikipedians in a list of article requests. Since it is not designed for indexing, vocabulary control is done rather rudimentarily in the form of redirects. Finally, similar to other universal encyclopedias, it recruits its vocabulary from more or less natural language currently in use, capturing consensual meanings of a term that have become relatively accepted.

WD, as a central storage repository, tries for the most part to manage WP’s data on
a global scale (Vrandečić/Krötzh 2014, 78) thus bringing encyclopedic and DL forms of KO together. Authority data also used in bibliographic databases are assigned to various types of entities which are part of WP articles. WD serves as a linking hub, allowing for relations of equivalence, and, only recently, for those of closeness and hierarchy between the WD item identified by an abstract identifier and the item in an external database identified for its part by a so-called external identifier. Special sitelinks connect an item to corresponding content on client wikis, such as WP. By mapping external entities, WD connects to more than 1,500 different sources of authority information. Moreover, if a source of authority information which is registered as external property has already been mapped to a source linked to WD, this prior mapping can be exploited for creating a preliminary mapping of the newly connected source to WD.

By providing structured data for the different WP language versions, variations of meaning between different languages are of particular relevance to WD. As other languages often come along with cultural specificities embedded in particular political, economic and social structures, a concept and the meaning of a term used to express it are language specific and cannot always be translated one-to-one into another language. These cultural foundations of KO that underline the fact that the web is not only a technical but also a social and cultural project (see García-Maro 2016) also need to be properly addressed.

5. Comparison

Clearly common to both forms of KO is that they stand for systematically organized forms of KO and deal with natural language. However, with regard to their most specific aims there are considerable differences between both. While WD items are grounded in an encyclopedic information architecture providing targeted access to factual information to help understand the meaning of a term, the prime motivation of a DL is knowledge representation and to enable information retrieval. Due to their relatedness to different kinds of IS with diverse inherent scopes, both forms of KO are characterized by different construction principles for concept modeling. The frames of concepts can vary enormously due to differences between lexical meaning in natural language on the one hand and their DL meaning on the other hand. Moreover, thesaurus concepts are characterized by representation functions on various levels.

6. Recommended mapping model and approach

Considering the differences in modeling topical concepts between both forms of KO, we would like to argue for a mapping model as depicted in ISO 25964-2 (2013, 78p.) for mappings between thesaurus and terminology: i.e. a selective mapping (see Fig. 1).
At the working level, however, “establishing concept mappings (...) should follow the same general methodology and practices as between two thesauri” (ibid., 81). Against this background, the extension of mapping relation types in WD by accepting ‘close match’, ‘broader match’ and ‘narrower match’ as additional properties (Neubert 2017) is clearly approved. Taking the stated differences in concept modeling into account, we would like to present criteria which could provide guidance on approaching a mapping process between topical thesaurus concepts and WD items and serve as a starting point for identifying suitable concept candidates of a thesaurus by paying special attention to scope, representation function and the semantic stableness of a thesaurus concept.

- As depicted above, the detection of subject-specific authority data from other DL in WD, which had already been mapped to the respective thesaurus, could be useful to identify promising candidates to be mapped.\(^6\)
- The allocation of a concept to a certain subthesaurus or, more generally, to a specific top term could hint at the degree of abstractedness, that is, the concreteness of a concept.
- With regard to the degree of specificity of a concept, its location within different subthesauri or different branches of a single subthesaurus could provide information as to what extent several meaning components could be included in a concept. Also referring to the degree of specificity of a concept, the problem of hypernyms and hyponyms could be dealt with by identifying underspecified synonyms within an equivalence class.
- Referring to the stableness of subject-specific vocabulary, it was mentioned that huge differences exist between different domains. With regard to the extent of changes within one thesaurus, dynamics in concept modeling in the past could be approached on various levels: on the subject category level (e.g. adding or merging two subject categories into one), the concept level (e.g. adding a new concept, represented by a preferred term), and on the terminological level (e.g. adding new synonyms).

\( ^{6} \) In WD the interactive mapping tool Mix’n’match could be used for this purpose. http://meta.wikimedia.org/wiki/Mix%27n%27match. Accessed 30 January 2018.

\( ^{7} \) Already being mapped to the subject headings of the GND it turned out that 2,034 of the 5,339 topical concepts of the STW are already transitively linked to Wikidata items via GND ID (Neubert 2017, 11). The intellectual evaluation of a random sample, however, reveals differences in concept modeling between both forms of KO as depicted here.
reversing the preference between a preferred and a non-preferred term). Hereby, it could be possible to identify particularly stable or unstable areas below the subthesaurus level within a thesaurus. The SKOS publication of the thesaurus could be used to track these changes (Neubert 2015).

The procedure depicted, meant as an initial stocktaking exercise, could help to select a core set of fairly stable topical thesaurus concepts whose degree of overlap between documentational and lexical meaning is rather high. The topical thesaurus concepts receive a fingerprint, so to speak, which expresses their suitability for mapping. Armed with this corpus of concepts, the editor is then able to begin with a systematic mapping process. Special attention must still be paid to the different language versions. The mapping in WD could be given an additional predicate by assigning a preferred rank, compared to levels of determinations as used in other mapping contexts.

7. Conclusion

To conclude: in a process which has already gained momentum, WD could serve as a core reference and linking infrastructure helping subject-specific authority data to break through the ‘container’ of closed collections and to “improve their relevance in the digital age” (Garciá-Marco 2016, 195). With regard to thesaurus maintenance and development, WD could serve as a core reference hub for concept modeling, helping to reconcile collection and web-scale areas of reference and thus leading to a greater harmonization of subject authority data. Thesaurus editors and contributors could immediately benefit from WP’s definitions and encyclopedic information; and WD’s data storage could “assist editorial tasks such as proposing and assessing intralingual and multilingual, preferred terms and equivalences” (ibid.).

Employing WD as a central hub for mapping could lead to a stronger convergence of conceptual modeling within the different sources of authority information, which should be undertaken as a constant endeavor. By bringing authority data from various sources together, it could assist interoperability and vocabulary sharing, because it helps to scrutinize the degree of conceptual overlapping. In a step-by-step process this could foster a new paradigm of standardization already on the rise which favors compatibility instead of homogeneity.
References


Phenomenon-based vs. disciplinary classification: possibilities for evaluating and for mapping

Abstract
The possibility of interdisciplinary research through online library catalogues is today a major challenge for knowledge organization. By examining the different approaches used to index documents in integrated library systems, interdisciplinary researchers may find that classifications based on academic disciplines are less appropriate for this type research when compared with phenomenon-based systems. Based on this consideration, the aim of the present paper was to compare two systems having different conceptual structures: the Dewey Decimal Classification (DDC), which is based on an epistemological approach, and the Integrative Levels Classification (ILC), based on an ontological one. We also explored how ILC phenomenon classes can be mapped to corresponding DDC disciplines and how DDC can be used in phenomenon-like ways. This work highlights the importance of future studies in phenomenon-based classification for the development of new indexing schemes and for more powerful applications of already existing disciplinary systems and opens a new way for improvement of interdisciplinary research.

1. Introduction
An important trend in recent research in knowledge organization (KO) is the criticism of academic disciplines as the main dividing principle of classification schemes (Beghtol 2010, 1056; Broughton 2015; Gnoli 2016). There have been many claims that disciplinary classifications may act as an obstacle to interdisciplinary research (Szostak et al. 2016), and may leave out such important applications as serious leisure, services for citizens looking for practical information associated to real-life events (Bonner 1982), particularly in e-government, or vendors' catalogues of such products as clothes and food, which do not fit in any academic discipline. However, phenomena seem to be a more general unit of knowledge than disciplines in ontological terms, so that general systems could be primarily defined in terms of phenomena, and a discipline could then be defined as a phenomenon belonging to the class of human culture.

Experimental phenomenon-based classification systems have been developed since the times of James Duff Brown's Subject Classification (Beghtol 2004). An in-depth enquiry towards a general classification of phenomena was developed by the Classification Research Group (CRG) (Austin 1969; CRG 1969) and, although this did not lead to any definitive system, two schemes were produced by CRG members, namely the Bliss Bibliographic Classification 2nd edition and the Broad System of Ordering, which do provide some main classes (3 and 088 respectively) under which phenomena should be classed along with classes for disciplines. CRG research has also inspired the Integrative Levels Classification (ILC) project, which has now been in
development for a decade and currently offers a general scheme with more than 8,000 classes and facets (Gnoli 2017).

On the other hand, the classification schemes used in most libraries and their online catalogues are still the classical disciplinary ones, like the Dewey Decimal Classification (DDC), the Universal Decimal Classification (UDC) or the Library of Congress Classification (LCC). This situation suggests that a comparative evaluation of phenomenon-based vs. disciplinary indexing, as well as an exploration of possible ways to create cross-references and mappings between the two kinds of systems, may have important implications for the future of classification. This paper explores such possibilities by specifically considering DDC as an example of a disciplinary system and ILC as an example of a phenomenon-based system.

2. Methods

Preliminary tests of comparison between DDC and ILC were performed by indexing small samples of items using both systems and considering how the resulting orderings and retrieval possibilities differ. An initial sample consisted of books on nature conservation held at the University of Pavia ecology library (Szostak et al. 2016, 104-106).

Experiments have recently continued by applying both DDC and ILC to items recorded in the Basel Register of Thesauri, Ontologies and Classifications (BARTOC) maintained by Basel University Library (Ledl and Gnoli, 2017). As a directory of thousands of knowledge organization systems (KOS), BARTOC covers the whole spectrum of knowledge fields, but does not require a very detailed indexing as the coverage of a KOS is usually quite broad. All BARTOC items, as they are input into the directory, are indexed by one or more three-digit DDC classes freely combined and by one or more EuroVoc terms. Additionally, using one or more ILC classes, also freely combined, we have indexed a KOS sample consisting of the first 200 Top-Rated Systems (that is the systems which have been voted as most relevant by BARTOC users) plus all the ten remaining KOSs dealing with health care; the latter having been chosen in order to test ILC within a specific domain of knowledge. To allow reference to a stable source, the first edition (ILC1) of ILC was used, despite a second edition (ILC2) currently being developed which includes various changes and improvements. Update of classmarks in BARTOC to ILC2 is envisaged in the medium term.

Further work has consisted in mapping between DDC and ILC main classes, by creating an additional "DDC map" field in the MySQL database of ILC2. DDC appears to be a suitable reference for mapping to a disciplinary system, given its worldwide use. For similar reasons, terms in the Nuovo Soggettario subject heading system, developed by the National Library of Florence and based on a structural conception not far from that of ILC, are also mapped to DDC.

At the time of writing this paper (January 2018), the field has been compiled for 25
ILC main classes labelled a to y and for 100 DDC subdivisions labelled 010 to 990. DDC numbers, taken from WebDewey, the updated current edition of the scheme, have been used in full, so that, for example, ILC f"molecules, chemical substances" has been mapped to DDC 541.22 "molecular structure (in theoretical chemistry)".

Availability of DDC classmarks in BARTOC can subsequently allow leverage of ILC-DDC equivalences to automatically produce additional ILC classmarks, although these are expected to be less accurate than the classmarks produced intellectually because (1) the DDC classes used are very broad, (2) the very conceptual structure of the two systems is different. Concerning the latter point, some considerations are developed in the following part of this paper.

3. How to map a phenomenon to a discipline

As a disciplinary system, DDC scatters a given phenomenon, such as "water", into a variety of different disciplinary classes (chemistry, hydrology, hydraulic engineering, architecture, …). This can easily be observed by checking a term for the phenomenon in the DDC Relative Index, where references are provided to its occurrence in the various classes. However, the phenomenon "water" cannot be separated from the final class numbers assigned to the information objects. On the other hand, in ILC, water has a place of unique definition (Farradane 1950) at febb under class f of chemical substances; it can then be combined freely with any other concept, e.g. wkf7febb "fortifications with water".

So, should febb be mapped to a DDC class in chemistry, in hydrology, in hydraulic engineering, or in architecture? Concepts like "water" have been described by a DDC editor as topics, which could indeed be useful for "identifying equivalence and hierarchical relationships between Relative Index headings, one of the steps required to fully implement a topic-based data model for the DDC" (Green 2014). Topics can thus be another relevant unit for classification, in some way complementary to disciplinary classes:

"As a knowledge organization system, Dewey is an analytico-synthetic classification system first and foremost organized by disciplines, in which topics and concepts are scattered throughout. What can be done with Dewey as a classification system (that is, because it groups topics into classes) that couldn’t be done if it were strictly topic/concept-based, like the typical thesaurus? […] Many of the topics within a class occur as Relative Index terms. Within WebDewey, these are formatted in Marc 21 for Authority data and include references. Can/should Relative Index terms be used (as supplements to subject headings or independently) in indexing?" (Green 2017)

Green's idea involves the establishment of topic-to-topic relationships, which could improve search in DDC applications, such as a library catalogue: users could thus navigate from a water-related class to another in a different discipline.

Although topics are not defined precisely and are labelled by a term other than phenomena, it seems that the two units can work in similar ways, at least functionally.
Thus, Green's suggestion opens up the possibility of using DDC in some phenomenon-like ways.

Among other things, topics can work as a better reference than disciplines in mapping DDC to different systems, including ILC itself or subject headings and thesauri, which usually index phenomena rather than disciplines. A search for "water" in the Web Dewey Relative Index yields many different classes, including 551.35 "water--geological agents", 354.36 "water--public administration", 714 "water--landscape architecture", etc. However, all these are preceded by a number for the topic "water" alone, 553.7, also known as its interdisciplinary number.

To be precise, even the number 553.7 is not discipline-neutral, as it belongs to the hierarchy of 550 "Earth sciences and geology", which is a discipline too; still it is indicated by the Relative Index as the most general number available for the topic "water". As such, 553.7 is supposed to be used to class interdisciplinary works, like a book dealing with water from many different perspectives. Many topics in the Relative Index are provided with such interdisciplinary numbers. These are also used to map the terms of Nuovo Soggettario to DDC (Lucarelli pers. com.). Farradane would say that the place of unique definition of water in DDC is Earth sciences. A phenomenon-based system would be ontologically more precise, claiming that the place of unique definition of water is terrestrial materials or substances, which can be studied typically by Earth sciences but also by a variety of other disciplines including chemistry, astronomy, engineering, architecture, etc. In the ILC schedules, the most typical disciplines which study a certain class of phenomena are indicated in square brackets, e.g.:

\[f\] "molecules, chemical substances, pure substances" [chemistry, molecular physics]

4. Comparing the structures of DDC and ILC

Our mapping work also provides interesting opportunities to compare the disciplinary system and the phenomenon-based one. For example, the order of main classes of phenomena in ILC2 can be compared with the order implied by the corresponding DDC numbers:

\[a\] forms, mathematical objects [formal sciences, logic, mathematics] \(\simeq\) DDC 510, 160
\[b\] spacetime 530.1
\[c\] branes [string theory, M theory] 539.7258
\[d\] energy, elementary particles, waves [quantum physics, high energy physics] 539.7
\[e\] atoms, chemical elements [atomic physics] 541
\[f\] molecules, chemical substances [chemistry, molecular physics] 540
\[g\] continuum bodies, bulk matter, states of matter [classical mechanics] 530.4, 530
\[h\] celestial bodies, cosmic structures [astronomy, astrophysics, cosmology] 520
\[i\] minerals and rocks [geology] 550
\[j\] land, landforms, regions [physical geography, topography, geomorphology] 910
\[k\] genes, genetic code [genetics] 576.5
\[l\] bacteria, prokaryote organisms [microbiology] 579.3, 570
\[m\] organisms, eukaryote organisms 570
Inversely, ILC classes (of which only a to j have currently been updated in ILC2) can be rearranged using the corresponding DDC notation:

000 ≈ V  general class, the universe of knowledge
004 ≈ wtu  computers [computer science, information technology]
010 ≈ y99x  indexes, bibliographies, catalogues
020 ≈ spgiWl  library and information services [library and information science]
030 ≈ y99y  dictionary, encyclopedia, directory
050 ≈ y99j  academic journal, research serial
060 ≈ so  organized civil society, NGOs [organizational studies]
070 ≈ y99e  newspapers
080 ≈ y976t  references [documentation]
090 ≈ y97  document instantiations, knowledge items
100 ≈ yy  philosophy, wisdom
110 ≈ aWz  world, universe [ontology, general philosophy]
120 ≈ p98  objects, phenomena, contents, qualia, representations
130 ≈ yg  magic knowledge
150 ≈ p  consciousness, subjects [psychology]
170 ≈ p938  values, sakes, goals, happiness
180 ≈ yy99  philosophical schools
190, 140 ≈ yy99m  modern Western philosophy
200 ≈ r  rituals, tradition, mores, spirituality, religion, sacred, holy
210 ≈ r99  ritual kinds
220 ≈ r97s  scriptures, religious texts
230/280 ≈ rt  Christianity
290 ≈ rX  faiths, doctrines
300 ≈ s  communities, civil society, groups, people [sociology]
310 ≈ s97  members
320 ≈ t  polities, political units, governments [political sciences]
330 ≈ u  economies, economic systems [economics]
340 ≈ t93  regulations, institutions, legal order, laws, norms, rules, acts [law]
350 ≈ t95  procedures, protocols, administration, management, power exercise
360 ≈ sp  public institutions, public bodies, services of general interest
370 ≈ spe  schools [didactics, pedagogy]
380 ≈ wn  vehicles, means of transport
390 ≈ r95  rites, practices, worships, uses, habits
400/410 ≈ q  languages, idioms [linguistics]
Several remarks can be made from examining these comparative schedules. An obvious one is the different arrangement of the two systems, already discussed by Ledl and Gnoli (2017): while ILC follows the order of integrative levels in nature, thus taking an ontological approach, DDC follows the traditional arrangement of disciplines of reason, of imagination, and of memory, taking an epistemological approach. The
difference in approach can be seen clearly in such classes as DDC 001 "knowledge" or 030 "dictionaries", which in a DDC epistemological view are premised to the whole of knowledge, but from the ILC ontological perspective are simply kinds of high-level cultural phenomena in class y.

Another consideration is that some main classes of phenomena in ILC, such as d "elementary particles and waves" or l "prokaryote organisms", are represented in DDC by several digits as they are only subclasses quite deep in the hierarchy; conversely, main DDC classes such as 800 "literature" or 410 "English language" are just subclasses (of x "arts" and qem "Germanic languages") in ILC. Indeed, the top level classes are very different between the two classification schemes.

A further point of interest is that some entries in ILC are mapped to more than one entry in DDC:

- 190, 140 ≈ yy99m modern Western philosophy
- 230/280 ≈ rt Christianity
- 400/410 ≈ q languages, idioms
- 510, 160 ≈ a forms, mathematical objects
- 660/680 ≈ vt industry
- 740, 760 ≈ xg drawings
- 800/890 ≈ xl literature
- 900, 940/990 ≈ t91 historical periods

5. Cross references by phenomena

Topics, as introduced above, are also an implicit basis for cross-references within the DDC (Green 2014). These can be recorded in WebDewey as "see-also" references, or in library systems as links between classes provided in interfaces for navigation.

The latter kind of relationships is, unfortunately, not common in library OPACs, but has been successfully implemented in the SciGator web-based browsing interface (Lardera et al. 2017). This application allows navigation between DDC main classes and a selection of their subclasses before launching a search by DDC notation in the University of Pavia online catalogue.

SciGator's most original feature is query expansion using cross-references between DDC classes or their mapping with local shelving schemes. Topics, and indirectly phenomena, are indeed an intuitive basis for the identification and establishment of these cross-references in SciGator. Behind links between fluid dynamics in physics, hydrology in Earth sciences and hydraulic engineering in applied sciences, which are created by librarians in Pavia for SciGator users, stands the concept of water as a phenomenon. Such cross referencing has the potential of being more formalized and connected with topics listed in the DDC Relative Index.
6. Conclusions

Our experience shows that, while both DDC and ILC may be used in a post-coordinate way to assign a plurality of classes to each knowledge item, e.g. to each KOS listed in BARTOC, the general ordering of items and the display of search results are very different if phenomena are considered as primary grouping criteria instead of disciplines. The same can be observed in the schedules of DDC and ILC main classes if arranged by one or the other system.

Work on mapping the two systems in more detail needs to be continued, as does indexing of BARTOC by ILC both intellectually and automatically, so as to make it more complete, accurate and updated as the development of ILC continues. Evaluation tests with a more formal methodology are planned.

Overall, the experiments we have performed using both disciplinary DDC and phenomenon-based ILC for indexing books (first test in Pavia), library catalogue items (SciGator) and KOS directory items (BARTOC) look promising. In particular, some notions implicit in existing disciplinary classifications, such as topics and interdisciplinary numbers in the Relative Index, could be analyzed and formalized more precisely to provide bridges towards phenomenon-based systems like ILC or Nuovo Soggettario.

These experiences suggest that continued research in phenomenon-based classification is an important requisite both for the development of innovative KOSs and for more powerful applications of existing disciplinary systems.

References


Diane Rasmussen Pennington, Laura Cagnazzo

Relationship status: libraries and linked data in Europe

Abstract
Considering the increasing attention that the discussion on ‘Semantic Web’ and ‘linked data’ has gained in the international scene over the past decade or so, we deemed it useful to undertake a study to explore the role that libraries are playing in the linked data context. Focusing the sample on European national libraries and Scottish libraries, we applied multiple data gathering methods and constant comparative analysis to analyse these circumstances. The findings showed a general lack of awareness of the Semantic Web and the potential opportunities for linked data. Through the collection of viewpoints from information professionals, it was possible to identify reasons behind implementation or lack of implementation, as well as perceived benefits, encountered challenges, and outstanding issues requiring further investigation. Finally, we have drawn recommendations as well as remaining concerns for institutions contemplating linked data implementation.

Introduction
Libraries shared metadata long before the advent of the Internet by utilising standards such as MARC and AACR to enable interoperability (Tallerás 2013). A key difficulty with their use in the networked environment has been the inability to exchange data at scale, due to the diversity of descriptive standards and schemas adopted as well as the diversity of languages in use (Breeding et al. 2016). Linked data offers libraries the means for enabling interoperability, improving data management, and enhancing the amount and quality of information available to more people (Byrne and Goddard 2010). Linked data has the potential to improve discoverability of library data, eventually helping libraries realise their dream of appearing within those coveted first ten results when searching the open Web. With linked data, people lacking knowledge of library jargon and metadata standards can finally benefit from the rich information stored in libraries’ catalogues and other online resources (Rasmussen Pennington 2016). After developing the WWW protocol, Sir Tim Berners-Lee, frustrated by the fact that it was exclusively made of human-readable content, introduced the idea of an extension of the Web enabling relationships not only between documents, but also between the data within documents (Tallerás 2013). That marked the origin of the Semantic Web: an advanced version of the traditional Web in which users make their own contributions, but they must express all content in a machine-readable, interoperable way (Berners-Lee 2000). If the Semantic Web is the overarching concept and the ultimate goal, linked data is the means to reach that goal (Bizer et al. 2009).

We are approaching the end of cataloguing records containing siloed library-provided data, and we are moving towards more enriched data sets originating from various resources. This requires library data to be structurally flexible and applicable
to multiple online contexts (Coyle 2009). Essentially, “We are moving from cataloguing to catalinking” (Wallis 2013, slide 19).

This presentation highlights the convergence of two related studies: linked data implementation across European national libraries (Cagnazzo 2017), and linked data awareness and use among Scottish libraries (Rasmussen Pennington 2017). Whilst the first offers a more global view, the latter provides a targeted lens on one country.

**Objectives**

This work aims to investigate the reasons for linked data adoption and non-adoption in Scottish libraries and European national libraries. Additionally, it seeks to understand the perceived benefits of and challenges surrounding adoption, based on the experience of information professionals who are directly involved in implementation projects. This will offer a comprehensive picture of how European national libraries and Scottish institutions perceive the call of the Semantic Web and how they are responding to it. It will present use cases to exemplify the different purposes linked data can serve, identify technical difficulties and outline solutions adopted to overcome them. Furthermore, it will report viewpoints on the pros and cons of linked data adoption and the motivations behind organisational implementation decisions. Finally, we will provide recommendations for institutions wanting to embrace the Semantic Web, including how to overcome common issues faced in library linked data implementation initiatives.

The research questions addressed were as follows:
- What are librarians’ understandings of the concepts “linked data” and “Semantic Web”?
- What are the perceived benefits of library linked data implementation?
- What are the reasons behind libraries’ decisions to implement or not implement linked data?
- What are the barriers to linked data adoption?
- What are the challenges encountered by libraries before and during linked data implementation?

**Methodology**

A qualitative case study approach provided an in-depth analysis of a few settings. To answer our research questions comprehensively, we utilised multiple data gathering techniques:
- **Literature review**: our starting point to investigate the status of linked data implementation and use across European national libraries.
- **Semi-structured interviews (n=15)**: we interviewed information professionals working with linked data or metadata in European national libraries. We chose interviews for their ability to expressing breadth of perspective and individual viewpoints. Interview data helped us integrate and enrich the information.
collected through the literature review.

- **Online resources analysis (n=26):** We used browser extensions able to recognise Semantic Web structures (such as Semantic Radar) to identify whether non-participating European national libraries’ websites had implemented linked data.

- **Survey (n=113):** Public, academic, school and national librarians in Scotland participated in an online survey aimed at understanding the awareness, perceptions and adoption of library linked data.

We analysed these data through constant comparative analysis, an analysis approach used on disparate resources. The method consists of comparing each piece of data with other data that are similar or different, formulating concepts on possible relationships among data, and building categories to group the data (Onwuegbuzie *et al.* 2012).

**Results**

The combination of the different data gathering techniques employed allowed the delivery of a comprehensive answer to the research questions. What emerged from this study is a strong need to increase Semantic Web awareness and its potential within libraries. Librarians are commonly uncertain about the meanings of “Semantic Web” and “linked data”. The case of Scotland is emblematic of the obstacles that most institutions experience in adhering to Semantic Web implementation: in fact, a limited percentage of libraries have been using linked data, although a growing number of governmental policies require the publishing of public information as open data.

Professionals expressed viewpoints on perceived and experienced benefits and challenges of linked data implementation at their institutions. The most common barriers hindering linked data implementation were lack of expertise, lack of time, lack of staff, licensing constraints and difficulties in obtaining management buy-in.

The online resources analysis afforded the identification of RDF structures on the National Library of Ukraine and the National Parliamentary Library of Georgia’s websites, as well as implementation efforts at the National Library of Switzerland.

**Conclusions**

A minority of participating institutions (n=9 national libraries; n=8 Scottish libraries) have applied linked data. Some projects demonstrated the advantages of augmenting the visibility and discoverability of library data, supporting interoperability, and overcoming obstacles such as linguistic barriers. However, several issues remain unsolved which demand further research, such as licensing constraints and reuse vs creation of ontologies. Even where open data policies are government-mandated, libraries expressed a need for successful implementation examples in order to obtain management buy-in. Collaboration among institutions is key, although not yet a universal reality, to enable further Semantic Web development.
References


Ricardo Eito-Brun

The Role of Knowledge Organization tools in open innovation platforms

Abstract
Open Innovation is defined as “the use of ideas and market knowledge – both internal and external – to develop innovations” (CEN/TS 16655-5:2014) and it is based on this premise: “Valuable ideas can come from inside or outside the company and can go to market from inside or outside the company as well.” Open Innovation has resulted in a set of collaborative tools to help involved agents develop networks where they can share ideas, exchange knowledge and propose new challenges, relying on what are called technology e-brokers or innovation markets, such as InnoCentive, NineSigma, Quirky, unseraller.de, TopCoder or Yet2.com. These tools consist of web-based platforms where companies can publish innovation challenges and other agents can analyze and propose potential solutions. They are used to capture external knowledge and to assemble it with the company’s internal knowledge base, acting as intermediaries where innovation-involved agents can establish links and start collaboration activities.

Although these collaborative innovation platforms or markets are extremely useful, their functionality could be improved with the help of knowledge organization system and terminology tools. The use of terminology tools to standardize these descriptions contributes to the localization of experts in specific areas. When posting a challenge, authors can make use of a closed list of terms to ensure the accuracy of the content. The demonstrator has been built with terms and descriptors from the National Library of Medicine Medical Subject Headings (MeSH) with the aim of showing the potential benefits of this approach and its technical feasibility.

In conclusion, it can be stated that terminology tools help improve access to data about partners, their skills and competences in complex information ecosystems that support Open Innovation processes and strategies.

1. Introduction
Innovation management is a sub-discipline of management that studies the rules that govern the generation, diffusion and adoption of innovation, and the relationships between innovation inputs and outputs. Innovation management was traditionally understood as a linear model comprising a sequence of activities – from basic research to serial production and market launch -, completed by a single entity. In the linear model, the achievement of innovations was determined by the planning, financing and execution of internal R&D activities or external technology acquisition.

Linear models were replaced by collaborative models based on feedback and interactions between different partners (Busse 2013; Harmelen 2012). Today, innovation management is seen as a non-linear, evolutionary, interactive process between the company and its environment that requires the close collaboration of different agents (Iordatii 2013). This evolution culminated in the Open Innovation model (Chesbrough 2003). Open Innovation states that valuable ideas may come from both inside and outside the company, and it comes as the result of knowledge specialization, availability of highly skilled workers, increasing capabilities of
suppliers, and the difficulties of having a complete domain consisting of all the aspects that need to be mastered in a successful innovation life cycle. Different entities or agents have a different level of participation in generating the knowledge streams that provide the inputs to create innovations (whether in the market, in scientific and technical knowledge, or social knowledge) and the complex interfaces between them.

The popularity of Open Innovation has led companies to consider the need for systematic planning of innovation and has extended interest in innovation management to companies – like SMEs – whose characteristics have made it difficult to set up an innovation management plan based on dedicated R&D departments and costly investments in the development of products or services with no guarantee of success.

Among the different manifestations of the open innovation paradigm and its adoption by a wide number of companies we found several web-based platforms where companies can post challenges to an open audience. Some of the platforms successfully deployed to support collaborative innovation processes include InnoCentive, NineSigma, Brightidea, InnovationExchange, Atizo, YourEncore, Battle of Concepts or Yet2.com, to name just a few. These platforms can be defined as “technical infrastructures for knowledge sharing, discovering and social interaction” (Bygstad 2010; Garcia-Barriocanal 2012).

Innovation platforms like those cited above, are mainly based on the Web 2.0 paradigm, offering companies the possibility of openly distributing and forecasting “business challenges” and collecting ideas from outsiders. In general, companies collect and analyze these ideas to identify potential improvements in existing products or requirements that may lead to the development of new products or services. The platforms offer functionalities to collect ideas and complete their assessment, with a clear focus on open, distributed collaboration. But the requirements of open innovation management processes may go beyond these Web 2.0 capabilities. In particular, there are two areas that are subject to improvement:

- Giving support to the staff in charge of analyzing and assessing the solutions proposed in response to challenges.
- Helping innovation managers make informed decisions on the feasibility of the solutions proposed by potential partners, and their actual capability.

In other words, once the answers to the posted challenges are collected and screened, innovation managers must take informed decisions on the feasibility of the proposed solutions. This is likely to depend on the internal capabilities of the persons, group or entity that proposed the solution. When moving forward with the innovation management process, information about the capabilities and experience of potential partners also becomes quite relevant, as decision makers must decide on the most appropriate ways to move from ideation to conceptualization. This step in the innovation management process will probably require additional assessments of the
potential partners’ capabilities.

This extended abstract presents the summary of an on-going study into the role of linked data technologies to complement web-based Open Innovation platforms. The research is aimed to provide an answer to the information needs of innovation analysts responsible for making decisions on the feasibility of innovations and potential partners.

The approach presented is considered particularly useful for SMEs, as the proposed solution can help them leverage their competitiveness and ability to innovate. This is due to two reasons: firstly, SMEs can easily promote and give visibility to their activities and achievements, in order to become partners of major companies; secondly, SMEs can find qualified partners to develop projects where they need complementary competences and skills.

2. The Open information ecosystem

Although not related to the Open Innovation concept, in recent years our society has also witnessed the emergence and progressive consolidation of a new approach to share knowledge, information and data, namely the Open Information model. This model was initially tied to the academic world. Academics began sharing reports and documents through web-based platforms known as open archives or digital repositories. The open information model was characterized as follows:

- An open publication model where authors can easily publish research results.
- An open access approach, as information and research results are freely available to the community, at no cost.
- Removal of the traditional barriers to accessing information: cost, difficult findability of information, etc.
- The development and homogeneous use of a set of technical protocols that automate data collection, aggregation and sharing (e.g., Open Archives Initiative – Protocol for Metadata Harvesting or RDF)

The development of this new paradigm to publish information has followed different steps:

a) Institutions launched their corporate open repositories, where their staff and employees could publish results of their research activities. This led to the existence of “information silos”, with a high number of distributed, local document-based repositories that were hard to exploit;

b) The development of the technical protocols cited above led to the creation of open databases or repositories that aggregated documents and document metadata harvested from local, institutional repositories. Both subject-specific and general-purpose databases became popular.

c) The initial focus of these repositories – centered on documents – developed a
broader outlook and repositories started to incorporate additional information about researchers, research groups and projects. The incorporation of data-sets into repositories, and not only documents, is also one of the results of this change in their initial scope.

This transformation has led to a situation where, instead of the initial term – “open access” -, it is possible to talk about “open information”. Today, academic repositories are moving from the document-centered “institutional repository” to CRIS (Current Research Information Systems) where data about research teams and their achievements is openly shared; Major initiatives like the OpenAire portal are becoming increasingly important to give access to results of research programs funded by the European Commission, like FP7 or Horizon2020. In fact, some of these projects request the participant companies to publish their project conclusions in open access, through these repositories (special clause 39)

These efforts are contributing to the creation of an open information infrastructure that – combined with the technical capabilities and protocols currently available for data exchange and aggregation – can be further explored beyond data or document retrieval needs. In fact, these platforms are called to play a significant role in the development and improvement of our current R&D and innovation capabilities, due to the following reasons:

a) Finding potential partners in both the academic or entrepreneurial world may become easier.

b) Innovation analysts can use the data available in these open information portals to assess the capabilities of potential partners for ideation or conceptualization activities.

c) Innovation analysts working with web-based open innovation platforms can use these portals to collect additional information about the researchers or research groups posting solutions to challenges; this background information may be extremely useful when assessing the quality and feasibility of the proposed ideas.

3. Information needs of innovation

To validate our hypothesis about the potential synergies between web-based open innovation platforms and open information infrastructures, the research team conducted interviews with R&D managers and analysts from Philips, Bombardier Transportation, AstraZeneca and Lilly. The interviews aimed to make an initial identification of the information that analysts consider useful when assessing the novelty and feasibility of ideas and the reliability of their partners. The interview findings confirmed the importance of seeking expertise in the problem area or in the technical solution proposed to solve the challenge at hand. Searching for expertise was considered one of the key activities when setting up partnerships and teams to work on challenges that will lead to an innovation.
The research conclusions enable identification of the following information items contributing to the assessment of ideas and partners:

- Company details, including lines of business and activities.
- Areas of knowledge, competences and skills.
- Description of the company facilities and resources.
- Description of projects (including product development) in which the company has participated.
- Entities and partners (universities, research centers, companies) the company has worked with in collaborative projects.
- References and clients, whenever possible with the context of the projects they have collaborated.
- Publications and data sets generated by the entity staff.
- Patents granted to the entity.

These data items will contribute to a resulting metadata infrastructure that can be used in two different ways in the context of a web-based innovation supported process: a) to help companies identify partnership opportunities in a global context with sound criteria, and b) to help companies assess the potential relevance of incoming ideas sent in response to “innovation challengers”, based on the level of confidence attached to the entity providing the idea.

The final information infrastructure incorporates metadata for recording and transmitting information about “business challenges”, although this point is satisfactorily covered by open innovation collaborative platforms. With this in mind, current functionalities of innovation platforms could be extended to incorporate “contracting opportunities” characterized by a set of objectives, target price, etc. In other words, the solution proposed in response to a challenge could be presented at a separate location where other entities could: a) start a process to find potential projects or challengers to create a team, and b) be identified as potential partners by other companies using the innovation platform.

The identification and analysis of these needs led to the development of a prototype platform that combines the capabilities of three different tools: a) web-based open innovation platforms, focused on the publication of challenges and solutions; b) business and expert directories, focused on providing contact details and basic administrative information; and c) content and data repositories (CRIS, open archives) where the skills and experience of staff and research groups can be known and assessed.

The proposed solution leverages the capabilities of standard business and expert directories:

- Business directories usually exclude SMEs, as they focus on larger companies or on companies working in a specific geographical area.
Information in these directories focuses on financial data and provides only generic activity codes that do not have the level of detail needed to establish partnerships based on companies’ knowledge, competences, skills and expertise.

Collaboration opportunities in the open innovation context should not be restricted to companies, but should include other agents like universities, academics, public research officers or knowledge and technology transfer agencies.

4. Envisioned solution and technologies

To validate the research hypothesis and analyze the feasibility of the envisioned approach, the research team completed the development of a prototype platform aimed to support the activities of innovation and R&D analysts. Specific requirements of the tool included:

– Maintaining a core dataset with information about experts and entities for a specific area of knowledge.
– Offering dynamic interfaces with portals in the open information infrastructure and patent databases to collect data about researchers, companies or research groups on demand.
– Offering interfaces to web-based open innovation platforms, so analysts working with them could easily interact with our repository.

The last point is significant, as the proposed tool does not aim to replace existing open innovation platforms but complement them with additional capabilities for their users. The proposed solution can be seen as an intermediate tool, midway between “open innovation platforms” and “open information infrastructures”.

The need for interfacing with different tools led to the adoption of the Semantic Web, or Web 3.0, technologies, making intensive use of Linked Open Data (LOD) technologies. If the current open innovation platforms are mainly based on Social Web, or Web 2.0 technologies, the automatic capture and forecasting of data and the automation of data exchange between platforms requires moving in the direction of LOD and the Semantic Web.

Activities that are the subject of data exchange automation include, among others:

a) identification of partners and collaborators in specific knowledge areas;

b) identification of opportunities or challenges suited to the skills and competences of the SME;

c) assessing the potential value of companies based on their previous experience and
d) assessing the potential quality of the ideas submitted by a company.

In the case of c) and d), the availability of company data in an easy-to-process format would enable the preliminary filtering of entities and ideas.

A decision was taken to focus the content of the prototype on the medical
engineering area. Following this decision, experts, research groups, companies and conferences were identified and data from them was collected and loaded, including their publications, projects and patents. A sample triple repository – supported by the Virtuoso tool – was built with data on 320 persons and entities. This repository can be searched using the SPARQL standard language, which makes it accessible from third party tools. Available access points include both persons and entity names, geographic location, areas of expertise, projects, and free text. SPARQL flexibility allows more complex queries, such as obtaining the list of persons related to a company involved in projects making use of specific technologies. To improve the relevance of the results, controlled vocabularies (MESH, Medical Subject Headings) have been incorporated to describe the competencies of persons and entities. Controlled vocabularies are in fact one of the main components of the proposed metadata infrastructure, since descriptions of businesses, technological challenges as well as skills and competences need to employ a common vocabulary to allow full automation of data searching tasks.

The second pillar of the technical solution is the ability to search for external sources of content (documents, projects, data sets of patents) related to a specific expert or entity. The feasibility of this functionality has been validated with interfaces to search for patents in the European Patent Office semantic data set, as well as projects, documents and data sets in OpenAire. Both repositories offer access to their data through the SPARQL query language, making it possible to build a data pipeline through web-based connectors and generate a consolidated view of the information distributed about a specific expert, company or research entity.

Controlled vocabularies play a decisive role in the implementation of the platform, and automatic assignment of LCSH keywords for describing entities, challenges and research capabilities has been incorporated through calls to the MeSHOnDemand tool.

5. Conclusions

Open innovation relies on the ability to seek and build partnerships with external experts and entities. In a knowledge-based, global economy, finding partners and having the data to make an informed assessment of their capabilities becomes a complex task. Existing web-based open innovation platforms focus on putting challengers and solvers in contact. The proposed solution is aimed to help innovation analysts in these complex tasks, by facilitating access to information that may help them make informed decisions on the feasibility of proposed solutions and the capabilities of potential partners. The solution acts as an intermediary tool positioned between web-based open innovation platforms and other information tools in the “open information infrastructure”, like open archives and repositories or open patent databases. The use of Semantic Web technologies makes it possible to streamline connections between all these tools and build a virtual space to explore the current data ecosystem in order to leverage companies’ innovation capabilities.
References


Abstract

This paper presents the results of a study on the topics that Mexican television stations assign to web-distributed news content, after seeing a resounding drop in web traffic indicators. The inductive-hypothetico-deductive method was chosen for the work: in 2015, a six-month diagnosis was carried out at 14 Mexican television stations; three to identify the indexing points used by the television stations and three to carry out a hypothetical plan, aimed at comparing results. Among the findings, it is clear that television stations use Content Management Systems to automatically assign basic metadata such as newscast title, date of the news item, broadcasting schedule and anchors. Other metadata in which topics are described are, however, incorporated haphazardly or insufficiently. The evaluation showed that new methodologies for analysis and documentary treatment of videos must be applied so as to refine content description and representation. Furthermore, the hypothetical plan exercise showed viability of increasing search engine relevance and with it, visibility, web positioning and access.

1. Introduction

Thematic metadata used in Search Engine Optimization (SEO) strategies from television news broadcasts published on the web may facilitate retrieval and increase visibility, as required by companies leading these business-model-based endeavors.

The prevalence of reduced user traffic within these audiovisual contents, discovered through search engines, led to this study. A possible relationship was seen between thematic representation, increase in visibility and access. This paper presents a review of 14 Mexican television stations and the way they apply thematic metadata to common source code SEO tags on the web.

The objective is to make use of all the linguistic and semantic capital generated in audiovisual journalistic production for representation and thematic retrieval. This would establish the foundation for improving search engine positioning performance and user access. “However, without any terms to associate with a multimedia document as with images or music, there is an inherent problem indexing such objects. It is possible that the document has some metadata associated with it, but this is not always the case (e.g., on the web). With the web, multimedia documents are becoming increasingly more readily available, and mechanisms to access such information are sorely required” (MacFarlane 2016: 181).

Web access to television contents by a greater volume of users depends, to some extent, on indexing by search engines, essential tools in enabling people to retrieve both complete programs and sections of them. (Sobak and Pharo 2017:739). In this regard, the objectives of human-computer interaction call for developing systems that improve yield and user satisfaction. (Tsakonas and Papatheodorou 2008: 1.234-1.250).
Therefore, content representation links metadata management and use to information retrieval. Considered intellectual activities and information science approaches, this makes linguistic reasoning central to information science (Engerer 2017: 661).

2. Methodology

The inductive-hypothetico-deductive method was chosen for this research on representation and thematic retrieval metadata exploitation in the sphere of Mexican television newscasts on the web and their effect on phenomena such as visibility and access. In this sense, it begins with the idea that the inductive-hypothetico-deductive spiral contains two essential procedural steps.

• Heuristic or discovery phase: a phase involving observation, description, reflection and inductive generalization, intended to generate hypotheses (which could actually be a solution to the problem, answer to the question or explanation of the phenomenon).

• Justification-confirmation phase: the process of verifying the basis of a hypothesis by means of a procedure or device designed to do so (and susceptible to being reproduced) (Sarabia 1999: 55).

The author thus proposes observation and description of the problem to be studied, in this case by analyzing how television news stations manage contents. Specifically, they carry out a diagnosis and, after exploring the situation, define an explanatory hypothesis regarding the behavior, causes and effects of the phenomenon, to model a situation that enables them to confirm whether there is thus an improvement in the problematic situation and to guarantee a solution for it.

Additionally, the research conducted rests on the idea of attending to the needs of dynamic users who consume diverse content on the web, including news content. “According to findings of a 2011 survey (Purcell 2011), 92% of American adult Internet users use search engines to find information on the web, with 59% who do so on a typical day. This and other studies confirm our intuitions regarding the important role of web information. The web continues to provide extremely low cost means of publishing information, often coupled with high incentives for doing so, since web content can affect purchasing behaviors, opinions, and other important decisions of web users” (Kakol, Nielek, and Wierzbieki 2017: 1.043).

3. Findings

The first step in the methodological plan was diagnosis. In this case, it had to be done externally, because, as in several industrial sectors, secrecy characterizes television broadcasting companies and upper management, offices and departments related to “first screen” television journalism and news production and “second screen” distribution and publication on the web. The analysis period was February-April 2015.

The study was set up once the phenomenon had been detected in February 2015,
when web traffic indicators went down 45% (due to the appearance of Netflix and other types of fixed programming) with respect to the previous year, both on web platforms and mobile web (web on mobile devices).

To establish a strategy allowing for more in-depth analysis of the problem of representation and thematic retrieval, research was conducted on data that appear on the web, following this first step.

The instruments used were: a) selection of television stations for the study, b) user interface analysis and c) source code analysis. Among the strategies defined, first of all a group was formed of Mexican television stations that broadcast news on the web as well as on television screens.

The Mexican television stations selected for the diagnosis were: Aprende Televisión Educativa, CNNE (in Spanish), Canal Once, Efekto TV, Excélsior TV (*), Foro TV/Televisa (*), Fuerza Informativa Azteca, Milenio Noticias (*), MVS Noticias, Noticieros Televisa, Proyecto 40, Telefórmula, Televisión Metropolitana Canal 22 and TV UNAM. Television stations that produce and broadcast at least one news program a day were taken into account. The cases marked with an asterisk (*) are television stations with journalistic subject matter that broadcast 24 hours. They all broadcast their normal programming on pay television and a web site (desktop or mobile). Regarding open television, this only applies for Canal Once, Foro TV/Televisa, Fuerza Informativa Azteca, Noticieros Televisa, Proyecto 40 and Televisión Metropolitana Canal 22.

The second instrument was user interface analysis, which consisted of review and analysis of web pages that publish television news broadcasts or any of their substructures (news, reports, interviews, chronicles, sketches, etc.). The purpose is to determine presence or absence of those metadata that represent the topic of the content exhibited and which are considered relevant for search engines to improve positioning and visibility and, therefore, retrievability. An inventory was made of the metadata to be identified in order to establish definitive metadata from the research.

The third diagnostic instrument was source code analysis, consisting of review and analysis of web pages that publish television news broadcasts or any of their substructures (news, stories, interviews, chronicles, sketches, etc.), with Firefox® as the browser through use of the inspection tool and reading of title tags, meta, h1, h2, h”n”, body and content, among others commonly used in SEO strategies.

For each television station or production entity, the news broadcasts available through streaming or on demand (VOD) were inspected, as well as five video clips corresponding to substructures (sections: national, international, sports, entertainment and culture or their equivalents). For this aspect, an inventory of the use of semantic topics to identify, was made

1. Density of key words
2. Tag title
3. Meta tags
4. H1 and H2 tags
5. Analysis of key words in the competence
6. Selection of key words by seasonality
7. Idiomatic variations and concatenations of key words
8. Tagging (internal and social)
9. Optimization of images
10. Meta-descriptions
11. Microdata use
12. Key words in URL

From the findings identified upon concluding the analysis of television news broadcast user interface and the selection of video clips corresponding to substructures (news items, reports, interviews, etc.), the results were:

- All the television stations analyzed have content managers prepared to automatically assign the following metadata:
  - Title of the newscast/news item (the only one directly related to the topic)
  - Date of the newscast/news item
  - Broadcasting schedule/Date and posting time
  - Anchors

- 34% of the television stations do not include a summary or brief news item about the subject of the newscast/news video.

- 86% of the television stations do not include key words on the web pages published by the newscast/news video. The same proportion of proper names of people or places are not included as tags representing the topic.

- 79% of the television stations do not include tags on related topics that could be of interest to the audience and thereby keep it navigating within the web site.

- 14% of television stations do not include social share tools.

- 21% of television stations do not offer their news broadcasts live or on demand: Aprende Televisión Educativa, TV UNAM (platform under construction that broadcasts continuously with a video player) and CNNE. In the case of the first two, their main function is not journalistic, unlike CNNE which is.

- 100% of the television stations lack publication scripts, step outlines or transcriptions linked to their television news broadcasts.

- Only 7% do not offer news broadcast fragments: the TV UNAM platform, broadcasting continuously with a video player, is under construction.

Source code inspection produced the following results:

- Only 14% of the cases presented texts with over 15% key word density, placing them within the range. However, none of the digital properties inspected
presented over 30% density. In other words, even when web editors do publish “journalistic summaries”, they lack sufficient semantic elements for a search engine to establish aboutness.

- In 71% of the television stations, the source code for the digital properties does not include key words in the tag <TITLE>.
- 57% of the television stations do not include statements that describe the subject of the digital properties where their news broadcasts or video clips of substructures are hosted. The remaining 43% that do so suffer, however, from incorporating journalistic texts rather than texts with search-engine-appropriate key word density.
- 100% of the television stations use <H1> tags (H1 is the HTML element generally used to identify the most important headline on a web page. Examples of H1 would be a page title, post title, product name). In other words, they all place at least one title in the digital property that hosts their news broadcasts or substructure video clips. However, that title (which is published in the user interface) does not always include at least one key word representing the topic. Furthermore, none of the television stations use <H2> tags to create synopses, subtitles or subsections in the corpus of “journalistic summaries”, thereby wasting an opportunity to improve key word density.
- In 100% of the cases analyzed, it can be deduced that television station web editors pass up the opportunity to study the key words used by their competitors. There are even cases in which key words are not planted in digital properties. And reduced interest in key word use can be detected when comparing the web edition of the same subject or news item on different television stations.
- In 100% of the cases analyzed, use of seasonally selected key words was non-existent. Pending confirmation through a subsequent study, this may be because at certain times of the year consumption of this type of journalistic product heavily reduces user traffic, especially during vacation periods. Web editors could, however, add key words related to the trending topics of the day, yet, given the minor use of tagging, this resource seems wasted.
- In 100% of the cases analyzed, web editors were seen to have trouble defining a strategy to establish or select key words based on the application of idiomatic variations and concatenations. Instead, they tend to copy <H1> tag titles, which is neither wholly appropriate nor effective in terms of search engines.
- 57% of television stations do tagging. However, we note the lack of a clear strategy for doing so based on the creation of lists, indices, vocabularies or taxonomies. Efekto TV and Televisa stand out for being more structured, as they use key words in a way that is closer to thematic indexing. None of the television stations have a social tagging tool, limiting or reducing the possibility of learning
what words the audience uses to search and represent the topics addressed in news broadcasts and their substructures.

- In only 21% of the cases was it found that fixed images or the first image of the video player are not optimized, or rather are not utilized as a mechanism for reinforcing the topic, or that relevance or concordance between the image and content topic is lacking.

- In 79% of the cases, meta-description uses (tag use <META = DESCRIPTION) was not identified as a tool to improve the thematic description of the content by means of a summary or abstract structured for search engines. It only appears with a greater degree of consistency in the digital properties of Televisa, Foro TV and Efecto TV news broadcasts. The latter case stands out because it uses a specific tag: <META = ABSTRACT>.

- 100% of the cases revealed an absence of microdata that search engines can promote as values added on to the relevance and potential attraction of digital properties, shown in snippets from the results pages of a particular consultation.

- Lastly, in 71% of the cases, key words are not included in URLs. In the cases analyzed, URLs are used with digital property numbers or with the brand of the production entity or the section, or with an operation indication such as "streaming", "VOD" and "live". In the case of Televisa and ForoTV news broadcasts, the URLs are especially long, as they incorporate the title. It would be more effective if URLs were to be used that included one or two key words, configuring the contents manager so that it carried this operation out automatically.

With the strategies and instruments used, it was possible to make a working hypothesis. If a model is applied or the labels filled out for title, topic, summary and social labeling, based on professional information criteria experienced in information indexing and retrieval, better positioning of the news is assured in search engines.

An intervention program was presented to the 14 television stations, which was accepted by five different newscasts that broadcast at different times. The intervention was applied to two newscasts, while the other three acted as a means of control. It was carried out from September through November 2015.

The intervention encompassed the following steps:

- Construction of a work group
- Discussion, design and implementation of a temporary policy for metadata use for thematic representation of journalistic content. The policy was based on the video analysis and documentary treatment model applied to the process of thematic representation of newscasts on the web.
- Training for web editors in managing and using metadata for thematic representation.
• Implementation of the metadata policy when publishing newscasts or substructures of them, based on the extraction of descriptors and key words, as with the documentary analysis of newscasts, as well as the use of other elements such as scripts, transcriptions, translations and subtitles.
• Automatized data collection (ComScore, Google Analytics, MyMetrix, Videolog) on user web traffic behavior by newscasts. Graphs are presented below showing the traffic curve (single browsers), where the greatest drop occurred in September and how, after implementing the intervention, it went back up.

![Unique Browsers (thousands)](image)

**Figure 1**

4. **Formulation of points learned**

   The field of information science clearly has opportunities to help enrich web content representation and thematic retrieval, in terms of research, formulating diagnostics and designing interventions for ongoing improvement and innovation. In the specific case of television newscasts on the web, an extra effort must be made. While there is a *corpus* of text that goes along with the multimedia resource, most of the enriched content is found within that element itself.

   • Automatizing the use of metadata through television station content managers, predefining the topics by thematic sections or subsections.

Exploiting the value of semantic aspects to represent and retrieve content on the web (Soto and Naumis 2014: 85).

Increasing content relevance for search engines, along with its visibility, positioning and access by applying SEO tools and better practices.

Facilitating content representation and thematic retrieval, attempting to make it as ideal as possible and fit with users’ linguistic production style (documentary languages that approach natural language and social tagging). In this regard: “The main challenge in this context is to predict the most suitable retrieval model for a given user query and to cover the semantic gap between user information needs and retrieval models” (Ayadi et al, 2017: 1324).

Improving the user experience for people entering the website.

Contributing to the productivity expressed in the web traffic indicators and, hence, to the television station’s return on investment.

Text, audio, video and animation form a discourse whose very format calls for more specialized documentary analysis, where specific aspects are considered, including viewing, content analysis, summary elaboration and construction of thematic descriptions with semantic perspective for search engine optimization (SEO), as well as (controlled and social) tagging.

Documentary analysis of television news broadcasting on the web also demands that the personnel responsible for it stay updated and professionalized in competences that overlap with the skills of the web editor profile, in particular, and professionals in the creative industries in general (Mietzner and Kamprath 2013: 590).

In other words, for the human resource, it involves designing and structuring audiovisual products, keeping in mind linguistic and semantic relevance for improving representation, retrieval, visibility and user experience.

5. Final considerations

1. From the perspective of the worldwide television industry, this is a time for redefining plans aimed at audiences that currently play an ever-greater role as active audiovisual content users on the web than as conventional television viewers.

2. On the one hand, the previously passive and receptive audience now behaves actively and creatively, turning TV viewers into users. Therefore, contents must now be produced to be technically set up to distribute on conventional television and on web platforms and video on demand (apps and opt) that require thematic metadata for algorithms of representation, retrieval and recommendation in accordance with user habits.
3. Given the technological convergence that enabled the overwhelming evolution achieved by Netflix, among others, and the crisis of traditional business models, it is essential to have strategies for SEO and its enrichment through content analysis and video documentary analysis, in the case of newscasts on the web.

4. Thus, if “content is king”, as global television leaders point out, and that is supposedly what audiences with screen independence follow, then it appears that Google is the ‘Caesar’ which controls the web empire. Much more than a search engine, this global company sets policies and develops algorithms that rate the relevance of the thematic representation of content, also conditioning retrieval and, therefore, visibility and access. This has implications from the perspective of the information society and the wider economy, a discussion that, for practical reasons, will not be pursued here.

5. Despite the previous point, at least in the Mexican case, we observe an absence of organizational information policies regarding representation and thematic retrieval. This is not a strength of Mexican television stations, as reflected by their business indicators. A serious dilemma exists, because web publication demands investments and expenditures that should still be subsidized. Television reaches millions of people, while web publication in Mexico does not, thus compromising return on investment. It is a difficult situation, because television stations cannot sidestep the web, as they would no longer be in the market.

6. Finally, there is a demand for information professionals with competences to respond to the challenges of representation and thematic retrieval. Web editors, just like their supervisors, lack these specialized skills. Consequently, we can expect that in fewer years than you might think, there will be a great demand for specialized human resources.

References


Marit Kristine Ådland, Marianne Lykke

Tags on healthcare information websites

Abstract
This paper explores tags and tagging behaviour on health information websites using an empirical, user-oriented, exploratory case study. Taggers and editors were interviewed about tags and tagging, while taggers solved tasks that included applying tags to a website. This qualitative data was combined with quantitative data; a transaction log analysis that included all tags applied to the site during the logging period. Results showed that taggers and editors disagreed both on the purpose of tags and what characterizes a good tag. Taggers applied tags to describe article content, explain articles, request information, and value article content. Some of these show that tags are not only not only topical descriptions, but communicative by intent. This result can potentially inform the design of tagging features.

Introduction
The purpose of this paper is to explore tags and tagging behaviour on a health information website: how “aboutness” is expressed, what facets are used, how the meaning or topical content of tags is related to document content, what vocabulary is used by taggers, and what opinions users have of tags. We chose Cancer.dk case for the study. This Danish information website is intended for cancer patients and their relatives. Knowledge of tagging behaviour will guide the editors of information websites when setting up tagging features. We find it interesting to study tags in a system where metadata is not as visible to users, as compared to other systems with tags, such as LibraryThing or Delicious. Specifically, we seek to answer the following research questions:

1. What characterizes tags on Cancer.dk?
2. How do the tags on Cancer.dk compare to other tagging systems?
3. How do users and editors perceive tags, and how do they intend to use them?
4. Do the views of users and editors correspond both with each other and the nature of tags?

Methods and study designs
This is a user-oriented (Järvelin and Ingwersen 2010) empirical, exploratory case study (Bryman 2012) in which we studied user interaction with tags, tagging, and information behaviour in general on a specific website, Cancer.dk. In this user-oriented study, the user is seen as a part of the system (Järvelin and Ingwersen 2010). With tags, the user also provides metadata, which makes user tagging behaviour crucial in understanding the entire system.

A combination of qualitative and quantitative methods provides a holistic view, which creates a complex picture of tagging behaviour on Cancer.dk (Bergman 2008).
The quantitative data (gained from log files), provides insight into what is going on and makes it possible to explore whether an identifiable phenomenon is frequent or not. We counted tags of various types and compared the number of tags in different categories. The quantitative data provided an overview and identified some patterns. No tags were excluded from the analysis. The qualitative data provided explanations for the behaviour.

Three studies were conducted:

1. **Tagger study**: The tagger study was conducted when the tagging feature was newly launched. Eight cancer patient participants, all aged from 39 to 73, solved tasks using the tagging feature. Five of them applied tags for the first time. Participants were tasked with finding information on Cancer.dk, then applying one or more tags to an article with the requested information. Example: “Find an article about psychological reactions to a cancer diagnosis. Apply one or more tags to it.” Each participant solved four tasks. Two of the tasks differed so that the participants solved three different set tasks. After the tasks, the participants filled out questionnaires and were interviewed. The interviews were semi-structured with the goal of finding out about the patients’ understanding, motivations, and opinions of tags and tagging, (including their purposes for applying tags). In addition to questionnaires and transcribed interviews, we identified the actual tags the participants applied in the transaction log (see Tag study below).

2. **Editor study**: We interviewed three Cancer.dk editors about their experience with tags and tagging. The goal was to find out about their opinions, then compare this to user opinion. The editors were interviewed in December 2012. The tagging feature had been on Cancer.dk for roughly one year, which ensured editors had experience with it at the time of being interviewed. The tagging feature was maintained by its editors and closely connected to Cancer.dk. Editors wrote articles and structured the site in collaboration with medical and social experts. Methodologically, these interviews were carried out the same way as the interviews with the taggers, except by phone; this allowed for limited possibility in observing non-verbal signals. These interviews were a follow-up to a longer project that the participants had been taking part in for over a year. The editors included criticism of both the tags and the tagging feature. We see this as an indication that the telephone interviews gave sufficient information about the editors’ opinions.

3. **Tag study**: A transaction log from the tagging feature gave the broadest data material. The first tag from an external IP was applied on November 30, 2011. The original plan was to leave the tagging feature unchanged for one year and log all activity. However, after a few months with the tagging feature live, the Danish Cancer Society worried about a high number of what they called ‘irrelevant tags’.
They believed that the field for applying tags was used as a search field by mistake. Thus, they changed the tagging feature in September 2012. The main difference, between before and after the change, was the visibility of the field for applying tags. Previously, the field for applying tags was open by default at the top of every article. Subsequently, it was hidden, and users needed to click one of the buttons in the tagging feature to apply a tag. In both cases, the buttons for tagging were visible on top of all articles both before and after the feature change.

**Results – 1 Tagger study**

The participants applied between 5 and 27 tags each, for an average of 12.25 tags. There were 86 unique tags, 98 tags in total. If we look at the unique combinations of user, tag, and URL, there were 92 such combinations.

There were few synonyms within the collection of tags from each tagger (Kipp and Campbell 2006). But, in many cases they applied tags that were synonyms to or grammatical variations of words in the article. Thus, the participants did have an understanding of the importance of synonyms in a system like this. During the interviews, they were also allowed to talk about this before being asked.

Our analysis indicated a connection between computer skills, an understanding of the tagging feature, and a focus on applying tags as topical descriptors.

*Topical description* was dominant when applying tags at Cancer.dk. Some of the taggers stated that they wanted tags to be exclusively topically descriptive. Participants who did not apply topical descriptive tags all agreed that such tags could be useful. A focus on subject description was often connected to a focus on finding information. To the participants, the topical tags did not have to describe the general topic of the article; it was sufficient that that it described a subsection or an aspect of its topic. Thus, topical tags did not equal subject headings. Their requirements were not as strict.

Six of eight participants mainly focused on tags as topical descriptions; most of their tags corresponded with this view. Examples of their subject description tags were *soja* (soy) and *antihormon* (anti-hormone), both of which applied to articles about these topics.

Other purposes found were: tags to explain the content, tags to evaluate articles, and tags to express requests for additional information (see below). These tags represented attempts to communicate with the system, its users, or editors. All the participants agreed that topical tags were good, but they did not agree on whether other types of tags added value to Cancer.dk.

*Explicative tags* explained an aspect of a word or a specific word within the content of an article. Two of the participants took this a step further and discussed how to apply tags solely to explain difficult wording or other aspects of articles. It was not easy to see this motive in the tags, as explanatory tags look like topical tags. But the willingness...
to apply tags that were synonyms to article words could be seen as an indication of explicative tagging. Despite the similarities between topical tags and explicative tags, the motivation was different. Topical tags described what was already in the article, while explicative tags sought to add an explanation and thus additional content to the site (Berendt and Hanser 2007; Ådland and Lykke 2012). None of the participants described practical challenges, like how to know what part of an article a certain tag explains. This would only be a problem if the article concerned similar or opposite concepts.

**Opinion tags** expressed an opinion on the content of articles. Two participants (4, 8) used tags to value the article content. The tags were: *fremragende* (excellent), *vigtigt* (important) and *god nyhed* (good news). This type of tag is well known from other systems, like Delicious (Golder and Huberman 2006).

*Tags applied to make requests* did not necessarily relate to the content of the article. Instead, this tag was applied to request information. One participant (3) stated that she had applied such tags. The purpose of this tag was opposite to topical description. Another participant (6) formulated the contrast between topical tags and tags that express a wish and said: “*To me it is negative to write about something it [the article] is not about*” (6). She did not want these kinds of tags in the system.

Regarding purposes for using tags, participants focused on topical tags, or at least tags that related to the same content as the article they were applied to. These tags were useful for browsing and searching. For some, good topical tags meant good credibility for the site. Thus, it was a challenge that a number of tags were topically misleading.

Both tags requesting more information and explicative tags could be seen as supplementary tags (Berendt and Hanser 2007; Ådland and Lykke 2012). They added something to the article that was not there before. If these tags were easier to distinguish from other tags, they could aid the editors in improving the site and be less likely to mislead users.

**Results – 2 Editor study**

Practical use of Cancer.dk differs between editors. Editor E1 said: “I never use the search field because I know where things are.” (E1). Editor E2, on the other hand, used the search field a lot when she looked for something, because she knew the site so well that she always knew what to search for. They had opposite behaviours but similar explanations; both made sense.

Editors E1 and E2 did not use the tags themselves. Again, they said this was because they knew the site so well and thus did not need tags. This was a sufficient explanation. But it is worth asking whether their general experience with tags at Cancer.dk may also be an explanation. The editors had not tried to identify explicative tags or tags that requested more information. These tags could inspire them to make changes on articles
according to the taggers’ suggestions. When asked, this was a new idea to them.

When editors applied tags, especially in the beginning of the tagging period, some of them applied words that were already in the text, while others did not. This seemed to be the same for the taggers; some taggers applied tags they had already found in the article text, while others saw this as useless and did not.

Editor opinions about the tagging feature changed during the project. At the project’s start, editor E1 was open and positive, and hoped that the tags would improve the search feature and users’ general ability to find information. Within a year’s experience, she felt that neither the feature nor the tags had fulfilled her expectations. Many tags did not describe the content of the articles. Editor E2 was more sceptical from the beginning. After the feature had been live for a year, she concluded that it did not fulfil its purpose. Editor E3 was the newest of the editors, so she could not report on her expectations before the tagging feature was launched. She liked the idea that users could index content, but from the tags she concluded that the users did not understand the tagging feature. The different purposes users had when they applied tags was a challenge for the editors, as users preferred tags that described the topical content of articles.

Some synonyms may be impossible for the editors to use in the text. If the editors write Coloncancer and colon cancer in the same text, one of them will look like a misspelling. So, if one is used in the text, and the other one is applied as a tag, both variations are covered. A mix of English and Danish is the same. Breast cancer or brystkræft may be used in the articles, but breastkræft or brystcancer are mixed compounds and less ideal. Yet, as some users use them and search for them, if these kinds of compounds are applied as tags, they may serve as lead-in terms; the text will not be disturbed by compound words that are incorrect because of the language mix.

A good tag gives a summary, or ‘essence’ of the text, according to editor E2, and it should be given from the user’s viewpoint, as an opposing view to editors and health professionals. However, the main point is the content description, so that tags can be used for search and retrieval. Her explanation was broader than just saying that tags are synonyms to article text words, but her examples confirmed that she included synonym tags in her summary-description of good tags. For the editors, a bad tag is one that does not describe the article. This means that a good, clear tag that expresses an opinion about the article or request for information is still a bad tag.

The editors expressed different opinions about how damaging bad tags are. Editor E1 saw them as noise, which can be a challenge in searching and browsing (though not necessarily). Editors E2 and E3 connected bad tags with confusion and even loss of credibility for both Cancer.dk and the Danish Cancer Society. This diversity was also observed among the taggers. Some found that bad tags gave bad credibility, while others did not see bad tags as a threat at all.
The different purposes users had when they applied tags was a challenge for the editors. They knew that it was difficult to distinguish between various purposes when looking at the tags only. It was hard to say whether a certain tag was a description of the article, a supplement, a request for more information, etc. This made it hard to use the tags for searching and browsing, as it was difficult to know what lay behind a tag: whether it meant that a given article was about the topic described in the tag, or that an article lacked the topic indicated in the tag.

From the interviews, it is our impression that, in some ways, the editors did not want tags, but a controlled vocabulary. This would fulfil some of the purposes that both editors and users had when they applied tags. A subject language that includes synonyms and possibly includes relations between terms (e.g. hierarchical) would provide the lead-in terms that users and editors need. Such a solution is, however, contradicted by the editors’ view that tags are essentially the user’s voice in the system. A controlled vocabulary can never replace this, which the editors were also clear about.

Results – 3 Tag study

After 13 months, there were 25,253 tags in the log. This number includes all tags, empty tags, repeated tags, and errors that the user seemed to correct by adding another tag, i.e. every time a user hit the tilføj-(submit)-button in the feature for applying tags. There were 8.4 tags per URL.

Table 1: Aboutness: Tag categories according to the relationship between tag and article content

<table>
<thead>
<tr>
<th>Aboutness categories</th>
<th>Internal tags</th>
<th></th>
<th></th>
<th>External tags</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number</td>
<td>%</td>
<td></td>
<td>Number</td>
<td>%</td>
<td></td>
</tr>
<tr>
<td>General topic</td>
<td>75</td>
<td>3.25</td>
<td></td>
<td>235</td>
<td>1.02</td>
<td></td>
</tr>
<tr>
<td>Exact topic</td>
<td>263</td>
<td>11.40</td>
<td></td>
<td>592</td>
<td>2.58</td>
<td></td>
</tr>
<tr>
<td>Exact aspect of topic</td>
<td>231</td>
<td>10.01</td>
<td></td>
<td>391</td>
<td>1.70</td>
<td></td>
</tr>
<tr>
<td>Topic mentioned</td>
<td>957</td>
<td>41.48</td>
<td></td>
<td>2,086</td>
<td>9.09</td>
<td></td>
</tr>
<tr>
<td>Indirectly related</td>
<td>15</td>
<td>0.65</td>
<td></td>
<td>195</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>Too general</td>
<td>40</td>
<td>1.73</td>
<td></td>
<td>109</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td>Too specific</td>
<td>90</td>
<td>3.90</td>
<td></td>
<td>2,507</td>
<td>10.93</td>
<td></td>
</tr>
<tr>
<td>No relation</td>
<td>210</td>
<td>9.10</td>
<td></td>
<td>9,271</td>
<td>40.40</td>
<td></td>
</tr>
<tr>
<td>Empty tags</td>
<td>19</td>
<td>0.82</td>
<td></td>
<td>3,241</td>
<td>14.12</td>
<td></td>
</tr>
<tr>
<td>Meaningless tags</td>
<td>10</td>
<td>0.43</td>
<td></td>
<td>2,663</td>
<td>11.61</td>
<td></td>
</tr>
<tr>
<td>Article was deleted</td>
<td>397</td>
<td>17.21</td>
<td></td>
<td>1,656</td>
<td>7.22</td>
<td></td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td>2,307</td>
<td>100</td>
<td></td>
<td>22,946</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

Summarized aboutness categories

| Related                   | 1,541         | 66.79|    | 3,499         | 15.24|    |
| Not related               | 340           | 14.73|    | 11,887        | 51.80|    |
| Empty/meaningless tags    | 426           | 18.46|    | 7,560         | 32.94|    |
| **Sum**                   | 2,307         | 100 |     | 22,946        | 100 |     |
We analysed all tags that had been applied during a year, dividing them into the following categories:

1. Internal and external tags – an indication of who applied the tag.
2. Lay or professional – does the tag belong to a lay or professional vocabulary?
3. Aboutness – the relationship between the aboutness of the tag and the aboutness of the article.
4. Topical facets – what is the tag about?

Internal and external taggers behaved differently. The internal taggers were from inside the organization behind Cancer.dk. It was easy to address them as a group and encourage them to apply tags. However, the group of external taggers was more stable. Their number of tags varied less during the logging period.

Visibility of tagging feature: When the tagging feature changed in September 2012, the visibility of the feature was the only change. This resulted in a dramatic decrease in the number of tags. There were variations in the number of tags during the entire logging period, but none as dramatic as this. What this demonstrates is that the visibility of the tagging feature strongly influences the number of tags.

Aboutness: Analysis of the log files showed how difficult it is to apply tags. The aboutness categorization reveals challenges in how tags relate to the topical content of an article, (see Table 1). Due to the nature of this material, it is not clear from the logs whether the users were aware of these difficulties. Some tags are good subject descriptions, while many others are not. Some express the needs and experiences of both users and patients, whereas others requested answers. Many more seemed to be misunderstandings and individual thoughts. Mixed together, the tags as a whole were difficult to use and difficult to judge. One possible explanation for the high number of tags that did not relate to the content of the articles could be that users mistook the tagging field for a search field. If so, this is an example of how efforts have to be taken to inform users about how to properly use a website.

Facets: The tag content was influenced by the content of Cancer.dk. Many tags repeated words from articles. The tag topics more or less covered the topics covered by Cancer.dk as a whole. Tags from internal taggers covered categories that were more diverse and described article content from various angles. Their tags showed a more even distribution of tag facets compared to those of external taggers. These results conflicted with the expectation that external users can add new viewpoints to the systems. External taggers, in fact, applied more tags that were unrelated to the content of articles.

The internal taggers were predominant when it came to tags about thoughts, feelings, and psychological symptoms. It seems that in this sense, internal taggers were better at choosing tags independently of Cancer.dk than the external users were. This conflicts
with the expectation that end users can add new viewpoints, while those within the system cannot.

Tagging is communication with the system. In an extended, narrow folksonomy like this one (Peters 2009; Vander Wahl 2005), there is no bookmarking feature and thus the users cannot use the tags solely for their own purposes. They cannot use tags to communicate with themselves. The communication is targeted towards the system, the editors, and the organization behind the system.

Conclusions

We found a disagreement between taggers and editors on what a good tag should be like and for what purposes users should apply tags. This explains some of the ‘wrong’, irrelevant and non-topical tags, which were intended as means of communicating with the system through tags, as opposed to describing the content. But misunderstandings and mistakes may also explain ‘wrong’ tags. Editors saw topical descriptive tags as most useful and found the tagging feature attracted too many tags that did not meet this criterion.

Our results can inform the design of tagging features; visibility is essential to attract tags, and the information surrounding tagging needs testing. The disagreement between our user groups can also inform tagging features: the tags applied within the system will be influenced by whoever has permission to apply tags.

Previous studies frequently extract tags for indexing or retrieval purposes. Our study, in which all tags are included, confirm these studies as productive. The communicative aspects of tags found in Cancer.dk indicate that taggers do not necessarily distinguish between tags in different systems. If we look at systems like Twitter, tags are communicative by intent; they add information to the tweet and do not necessarily cover the topical content of the tweet. However, when moved to an information website, this behaviour is unwelcome.

References


Towards integrated systems for KOS management, mapping, and access: Coli-conc and its collaborative computer-assisted KOS mapping tool Cocoda

Abstract
Increased use of data sharing makes interoperability between Knowledge Organization Systems (KOS) ever more important, but concordances between these systems are rather rare. Project Coli-conc aims to address this gap by developing tools, methods and techniques to simplify and accelerate the intellectual creation of concordances. It also aims to ease their use and exchange and at the same time to provide quality monitoring that aids quality management. The project creates a set of reusable software modules to enable uniform access to KOSs, concordances and concordance assessments. These modules are provided as a web application to support effective processing of concordances. In addition, existing software has been evaluated and enhanced with new components for storage of, access to and analysis of different concordances.

1. Introduction
Coli-conc, under development, enables interoperability between Knowledge Organization Systems (KOS) with an initial focus on German library KOS. We describe the functionality and architecture of Coli-conc and its mapping tool Cocoda and develop a broader design vision.

Coli-conc will provide easy access to many KOS and to mappings between KOS. It fills a void by developing a technical infrastructure for managing many KOS and KOS mappings and keeping them current. The specific objectives of the project are:

- Develop a collection of freely combinable open source APIs and provide web-based services for uniform access to KOS and KOS mappings and to support computer-assisted intellectual mapping among many KOS in multiple languages.
- Provide guidance for best practices for quality assurance of KOS and mappings.
- Collect, store, and manage KOS data and mappings in a uniform structure.
- Support using these data for many purposes, including:
  - knowledge base for automatic subject cataloging (e.g. Petrus 2011) and assistance in manual subject cataloging;
  - knowledge base for retrieval through browsing hierarchies and/or automated or computer-assisted query expansion, adding synonyms and related terms;
  - mapping and enriching subject cataloging data and queries between systems.

2. Coli-conc system description (Fig. 1)
The system architecture emphasizes synergy between the components, a common infrastructure in which modules providing a given functionality (such as intuitive hierarchy browsing) as well as data can be shared across, reducing the effort for providing functionalities and data as compared to stand-alone implementation.
Fig 1: Coli-conc Functional Architecture

The key elements of the infrastructure are shown here and elaborated in the remainder of Section 2.
2.1. The KOS and KOS mappings database and the JSKOS data format

At the heart of Coli-conc is a database of KOS and KOS mapping data, including

- mappings imported from existing concordances produced by various mapping projects, such as KoMoHe (Mayr and Petras, 2008) and Wikidata (a linking hub to a large number of authority files and other KOS, Neubert 2017), and
- new mappings created in the Coli-conc project.

KOS and mapping data are structured using the JSKOS format to be discussed next. For each piece of data the sources (imported KOS and concordances, editors) are given.

Figure 2: KOS and KOS mapping data structured using JSKOS

```
{
  "type": ["http://www.w3.org/2004/02/skos/core#narrowMatch"],
  "fromScheme": {"uri":"http://dewey.info/scheme/edition/e22"},
  "toScheme": {"uri":"http://bartoc.org/en/node/454"},
  "from": [
    {"memberSet": [
      {"uri": "http://dewey.info/class/387/e22/",
       "notation": ["387"],
       "preflabel": { "en": "Water, air, space transportation" }
    ]}
  ],
  "to": [
    {"memberSet": [
      {"uri": "http://id.loc.gov/authorities/subjects/sh85121579",
       "preflabel": { "en": "Shipping" }
    ]}
  ]
}
```

We developed the JSKOS format for the unified representation of KOS and KOS mapping data from disparate sources (Voß, Leld and Balakrishnan 2016; Voß 2017). JSKOS is based on SKOS and JSON-LD. It allows conversion to and from RDF and to and from the MARC 21 Format for Classification Data (Heggø 2017). JSKOS was developed primarily for web applications, but it can also be used as the internal format of NoSQL databases such as MongoDB. JSKOS combines elements of SKOS and Dublin Core such as concepts, concept schemes, modification times, and publishers and extends these existing standards to cover a richer set of KOS and KOS mapping data, including the following:

- strict definition of how to encode repeatable and non-repeatable fields;
- treating mappings between KOS elements as first-class objects;
- confidence level for mappings;
- elements for concept occurrences and co-occurrences;
- an extensible list of relationship types beyond just USE/UF, NT/BT, and RT;
- mappings with multiple concepts, ordered lists;
- closed world statements.
2.2. The KOS and KOS concordance registry

This registry contains metadata for all KOS and KOS concordances that were or will be ingested into the Coli-conc database. Metadata include indicators for the quality of the data in KOS and KOS concordances; these quality indicators inherit to the specific data items in the Coli-conc database. After a review of KOS registries (Voß, Agne, Balakrishnan and Akter 2016), we focused on BARTOC (Basel Register of Thesauri, Ontologies & Classifications) and used its metadata schema as the basis for the more complete Coli-conc registry schema (Voß, Ledl and Balakrishnan 2016), which also refines the NKOS KOS types used in BARTOC with KOS types from Wikidata (Voß 2016), to be extended to a faceted classification of KOS types. The Coli-conc KOS registry contains a subset of BARTOC. It supports keyword search and filtering by KOS type and downloading KOS metadata.

2.3. Input processes

These include general data checking and cleaning, transformation to the JSKOS format, and ingestion to the database with detection and consolidation of duplicates.

2.4. The Cocoda mapping tool

The core part of Coli-conc is the mapping tool Cocoda, a platform for computer-assisted intellectual mapping. The main modules of Cocoda are:

- The suggestion tool provides mapping suggestions from several sources:
  - mappings from various sources/projects, including Wikidata, stored in the Coli-conc database; a source may create mappings manually and/or automatically;
  - implicit mappings derived from correlating subject descriptors (subject headings or classes) assigned to the same document in a library or union catalog (Buckland et al. 1999).
- The edit and collaboration module supports computer-assisted editing for high-quality mappings. An editor works with mapping suggestions and selects the best one, edits a mapping, or enter an entirely new mapping. The editor can also explore the source and target KOS to better understand concepts to verify a suggested match or find a better one. The collaboration function allows several editors to divide work on mappings for one KOS pair and communicate about specific mappings, and it supports expert review of the mappings.
- The KOS Representation Module supports KOS exploration through searching, browsing, and displaying the hierarchical and other KOS structure (see 2.5).
- The measure module for monitoring quality. It produces statistics on the types of KOS, frequency of usage on KOS and terms for search, Information on mappings: per KOS, per topic or subject field, their frequency, new entries, status of on-going mapping work.
Fig 3: User Interface - Mapping Tool Cocoda
2.5. User exploration interface

The KOS exploration interface is provided through a module used throughout the system, in Cocoda, in systems to assist catalogers, and in the end-user interface. It supports searching, meaningful display of hierarchies and concept maps (visual displays of concept relationships) for browsing, display of all information about a concept, and navigation through hyperlinks.

2.6. Terminology services and connection to cataloging systems

Convenient access to actual KOS data notations, concepts, terms, and relationships requires web services. A few systems make KOS data available via APIs (Voß, Agne, Balakrishnan and Akter 2016), but coverage and quality of terminology services differs widely. Queries to a terminology service should support applications such as browsing and searching in KOS in a uniform way. Based on a review of existing KOS APIs, we are developing a JSKOS API both as a wrapper for existing terminology services and as an independent terminology service (Dührkoph 2017). Access honors restrictions, such as those imposed by OCLC for the Dewey Decimal Classification.

As we mentioned above, the KOS and KOS mapping data in the Coli-conc database are very useful to catalogers. Cataloging systems can make these data available by importing them using the terminology service (under consideration by the semi-automatic subject cataloging tool Digitaler Assistant, Hinrichs et al. 2016), or they can arrange with Coli-conc for integrating use of Cocoda into their system.

3. Next steps

In the short term the project will focus on
- ingesting more data and making the system operational;
- improving browsing functionality and usability;
- defining criteria for mapping relationship types;
- moving to a completely integrated DBMS to ensure higher performance;
- automatic enrichment of VZG’s “Gemeinsamer Verbundkatalog” (union catalog).

In the medium term, we are thinking about
- extension to other KOS (museums, archives, all areas);
- support for KOS in multiple languages using existing knowledge sources, including dictionaries, WordNet, and multilingual KOS such as the Library of Congress Subject Headings (LCSH), the Universal Decimal Classification (UDC), the Dewey Decimal Classification (DDC), and the Unified Medical Language System (UMLS), which includes the Medical Subject Headings
(MeSH) among its 150+ KOS;

- making many automated mapping tools available (See Fifty …);
- implement measures for quality assessment.

4. A vision of Knowledge Organization System management utopia

In our utopia the functionality of a system like Coli-conc would be expanded to include the functionality of a full-fledged KOS management tool that combines the best features of existing tools, including deriving concepts, terms, and relationships from a corpus of relevant texts, creating and maintaining a meaningful hierarchical arrangement of concepts, an efficient user interface; and collaborative KOS management. The Coli-conc environment is ideal for creating and maintaining KOS, since an initial collection of concepts and terms and many kinds of relationships can be extracted from the KOS and KOS mappings database.

Both mapping and establishing concept relationships within a KOS can be made easier and better through a concept hub. (Fig. 4) (Soergel 2011). In the hub, concepts are expressed through combinations of elemental concepts, more precisely through description logic (DL) formulas (Bechhofer and Goble 2001). Two terms with the same DL formula, are likely to designate the same concept. If they are not, we need to refine the system of elemental concepts and/or the system of syntagmatic relationships used in the DL. If the two terms come from different KOS, the hub establishes a mapping; if they come from the same KOS, the two terms are synonyms.

The beauty of this approach is that a reasoner can infer relationships between DL formulas and therefore between the corresponding concepts. This can be used to assist in the elaboration of the structure of one KOS and in establishing mapping relationships other than equality between two KOS.

We hope to try out the concept hub approach in a pilot, comparing inferred relationships with verified relationships in the database.
5. Conclusions
We have presented a vision of an integrated systems for KOS management, mapping, and access and pointers towards the incremental and modular implementation of the functionality of such a system. The ideas presented should be helpful to designers and developers. Such an integrated system opens many possibilities for improving information systems to better support users.

Acknowledgements
Coli-conc is funded by the Deutsche Forschungsgemeinschaft (DFG) (German Research Foundation) grant no. GZ-DI 364/8-1 / AOB J 622525, extension GZ-DI 1364/8-3 / AOB J 642034 to the Verbundzentrale des GBV (VZG.), the Head Office of the Gemeinsame Bibliotheksverbund (Common Library Network)

References


Melissa P. Resnick, Elena Cardillo, Marc Jamoulle, Magdala de Araujo Novaes, Frank S. Shamenek

Towards the semantic annotation and the prevention of the loss of information of second opinion requests from rural Brazilian primary healthcare providers: the Q-codes use case – a work in progress

Abstract
Objectives. To support documentation, various terminologies have been created to assist in this activity. Only a few terminologies cover the General Practice / Family Medicine (GP/FM) domain (e.g., the International Classification of Primary Care - ICPC). As ICPC fails to capture some non-clinical issues (e.g., organizational and managerial aspects), the Q-Codes taxonomy has been developed to extend ICPC, encompassing those contextual professional issues. The aim of this work is to show the value of Q-Codes in preventing loss of information through the semantic annotation of Second Opinion Requests of rural Brazilian primary healthcare providers.

Methods. Question-answer pairs for the years 2010-2012, in the Brazilian-Portuguese language, was obtained from an urban telehealth center. Each selected question was read to determine its semantic meaning, and coded using both the ICPC and Q-Codes classification systems. Based on this meaning, each question was manually assigned between 0 and 5 Q-Codes.

Results. The majority of Q-Code assignments were almost equally split between the Patient’s Category”, QC (42%) and the "Family Doctor's Issue", QD (37%) domains. Domain QT, that is "Knowledge Management", covered 21% of the assignments, while a single assignment was made to the domain QP, “Patient Issue”. Six of the top 10 Q-Codes assigned belong to the cited QC domain, 3 to the QD, and 1 to the QT domain.

Discussion. Analysis demonstrates that lost information represents age (QC), health prevention (QD), and medical education (QT). Medical education is one of the reasons that the telehealth system was implement, while age and health prevention are important to providing healthcare in Brazil.

Conclusions. Preliminary results show that Q-Codes capture information that otherwise would be lost in the case of using only clinical coding systems such as ICPC.

1. Introduction
Medicine is a broad field with many specialties. To support documentation, storage, and retrieval of information, various terminologies have been created to assist with these activities. These include, but are not limited to, the International Classification of Diseases (ICD) (Lagasse et al. 2001), the Medical Subject Headings (MeSH) of the National Library of Medicine (Nelson 2009), and SNOMED-CT nomenclature (Wang et al. 2008). The majority of specialized medical domains have proper nomenclatures and classifications, mostly targeted at clinical and specialized medicine (Cornet and de Keizer 2008), thus only partially covering the General Practice / Family Medicine (GP/FM) domain. With the use of primary care related classifications, for example the International Classification of Primary Care (ICPC), it is still difficult for General Practitioners (GPs) or Family Doctors to target and code, in their daily practice, the non-clinical issues or contextual information (e.g., organizational and managerial
aspects) with the available coding systems.

To this aim, during the last few years the Q-Codes taxonomy has been proposed (Jamoulle et al. 2017), providing an extension of ICPC concepts for contextual issues, and focusing on the semantic aspects thus constructing a true semantic resource to be used as an indexing system.

In this paper, the feasibility and usefulness of Q-Codes is tested to provide a semantic annotation (using both ICPC-2 for clinical concepts, and Q-Codes for both non-overlapping clinical concepts and non-clinical concepts) of questions from Question-Answer pairs of rural Brazilian healthcare providers, as they seek, and receive, second opinions from urban telehealth centers. The aim of this work is to show the value of Q-Codes in preventing loss of information through the semantic annotation of Second Opinion Requests (SOR) of rural Brazilian primary healthcare providers.

2. Background

2.1. International Classification of primary care (ICPC)

The International Classification of Primary Care (ICPC) is a classification initially used for data retrieval in primary care (Boot and Meijman 2010). By 1987, ICPC was introduced by the WONCA International Classification Committee (WICC) (Soler et al. 2008; I. Okkes et al. 2000), and the second release, including additional inclusion/exclusion notes, was published in 1998. This release is updated online once per year, and the last international update (ICPC-2e-v.6.0) was published in April 2017 (World Organization of Family Doctors (Wonca) and Wonca International Classification Committee (WICC) 2017). GP/FM is very broad in scope, encompassing both clinical and contextual issues (Jamoulle et al. 2017). Clinical issues pertain to signs and symptoms, reasons for encounter, processes and diagnoses, which are covered by ICPC (I. Okkes et al. 2000).

As noted by the World Health Organization (WHO) (Who 2017), ICPC has:

a biaxial structure and consists of 17 chapters, each divided into 7 components dealing with symptoms and complaints (comp. 1), diagnostic, screening and preventive procedures (comp. 2), medication, treatment and procedures (comp. 3), test results (comp. 4), administrative (comp. 5), referrals and other reasons for encounter (comp. 6) and diseases (comp. 7).

This classification has been used for structured documentation of episode-oriented care in primary care since the 1980’s (Lamberts and Hofmans-Okkes 1996), and is now considered as a de facto standard in Primary Health Care.

However, ICPC only offers a partial solution as it covers only the clinical issues of GP/FM (Boot and Meijman 2010; Soler et al. 2008).

Over the years, extensions to ICPC for nutritional advice (van Binsbergen and Drenthen 1999), procedures (I. M. Okkes, Veldhuis, and Lamberts 2002), community
pharmacy (van Mil, Brenninkmeijer, and Tromp 1998), and chiropractic medicine (Testern, Hestbæk, and French 2015; Charity et al. 2013) have been developed.

As ICPC fails to capture some non-clinical issues, which are predominately organizational and managerial aspects of GP/FM, a newly developed hierarchical resource, called “Q-Codes”, has been developed to extend ICPC encompassing those contextual professional issues. The letter Q was used as it was available for use as a chapter in ICPC.

2.2. Q-Codes

The development of the Q-Codes taxonomy started from the 1987 work of Henk Lamberts, a Dutch professor in General Practice (Amsterdam University), who designed the system for topographic archiving of copies of retrieved articles in the documentation system of the research department of General Practice. It was a simple one level classification with 7 main domain categories. In 2007, MJ, a co-author of ICPC (I. Okkes et al. 2000), undertook to revise and develop the Q-Codes as a full extension of ICPC for contextual issues, focusing on the semantic aspects and constructing a true semantic resource to be used as an indexing system for grey literature (Jamoulle et al. 2017). The Q-Codes taxonomy consists of 182 terms, distributed among 8 domains (Jamoulle et al. 2017), each containing between 2 and 4 levels of granularity, which represents respectively 44 subcategories, 109 sub-subcategories, and 21 sub-sub-subcategories (Jamoulle 2016). The 8 domains include: “Patient’s Category”, “Family Doctor’s Issue”, “Medical Ethics”, “Planetary Health”, “Patient Issue”, “Research”, “Structure of Practice” and “Knowledge Management” (Jamoulle and Resnick 2016).

The “Patient’s Category” domain (QC) represents such concepts as age, gender issues, and social issues. The “Family Doctor’s Issue” domain (QD) describes communication, clinical prevention, and medico legal issues. The “Medical Ethics” domain (QE) covers bioethics, professional ethics, and infoethics. The “Planetary Health” domain (QH) deals with such areas as environmental health, biological hazards, and nuclear hazards. The “Patient’s Issue” domain (QP) includes patient safety, patient centeredness, and quality of healthcare. The “Research” domain (QR) describes research methods, research tools and epidemiology of primary care. The “Structure of Practice” domain (QS) covers such topics as primary care setting, primary care provider, and practice relationship. Finally, the “Knowledge Management” domain (QT) deals with teaching, training and knowledge dissemination (Jamoulle et al. 2017; Jamoulle and Resnick 2016).

Representing non-clinical issues, the Q-Codes taxonomy provides a resource to facilitate access to GP/FM information. The first aim of Q-Codes is its use as an indexing system for grey literature (Jamoulle et al. 2017), and e-learning applications.
Constructed on the basis of Semantic web technologies, Q-Codes could be considered as a lightweight ontology ready to be used in the semantic web domain, to be extracted in Web Ontology Language (OWL). We will describe, as a use case for the application of Q-Codes, the Second Opinion Requests (SOR) from rural healthcare settings in the state of Pernambuco, Brazil. One source of information is SOR from rural healthcare teams in Brazil (Resnick et al. 2013).

2.3. Second Opinion Requests

Rural healthcare teams (physicians, nurses, lay community health workers) provide basic care to those living in their area (Haddad et al. 2015; Sanches et al. 2012). Sometimes, however, the healthcare teams need to send their patients to urban areas for a second opinion (consultation) or to see a specialist, often requiring a great deal of travel at considerable costs (Alkmim et al. 2012). Lack of expertise amongst health professionals in the primary care sector, unnecessary referrals, and the difficulty of facilitating consultations with specialists led to the development of the Brazilian telehealth program (Joshi et al. 2011).

In 1999 the first module of the telehealth platform called HealthNet was set up at Nucleo de Telessaude (NUTES), which is located within the Clinical Hospital of the Federal University of Pernambuco (UFPE) in Recife, Brazil (Barbosa, de A Novaes, and de Vasconcelos 2003). By 2004, HealthNet was implemented and operational (de Araújo Novaes et al. 2005). The telehealth service, through HealthNet at NUTES, provides medical second opinions to the healthcare teams in the rural areas of Pernambuco.

When a rural healthcare team needs a second opinion, in order to provide care to a patient, they send their questions through HealthNet to the nurses and physicians at NUTES. The appropriate health professional provides a second opinion or an answer through HealthNet back to the rural healthcare team. These questions and their corresponding answers (question-answer pairs) are collected for data sharing and reuse.

Managers of these telehealth programs need a way to evaluate and plan interventions, which will, in turn, improve access to telehealth services. One way that this can be done is through the use of information classifications, like the Q-Codes. The question-answer pairs from the telehealth service at NUTES will be used for this study.

3. Methodology

A data set containing 5,580 question-answer pairs for the years 2010-2012, in the Brazilian-Portuguese language, was obtained from an urban telehealth center. Webinars and tele-ECG Q/A pairs were eliminated, giving 1,669. Among these, 550 questions (~33% from each of the three years) were randomly selected and deidentified for inclusion into the sample data set.
Each selected question was read by the first author to determine its semantic meaning, and coded using both the ICPC and Q-Codes classification systems. Based on this meaning, some general guidelines and the definitions of individual Q-Codes, each question was manually assigned between 0 and 5 Q-Codes.

When the question provided an age of the patient, the appropriate age group was assigned from the "Patient's Category" domain (QC). When the question pertained to gender issues, such as pregnancy or birth control, it was assigned appropriate concepts from the "Patient's Category" domain (QC). When the question represented a need for information not referring to a specific patient, it was assigned a concept from the "Knowledge Management" domain (QT). Finally, if the question represented disease prevention and multimorbidity, it was assigned the appropriate concepts from the "Family Doctor's Issue" domain (QD).

4. Preliminary results

As of the writing of this paper, 100 (18%) of the 550 questions from the sample data set have been attempted to be semantically annotated with Q-Codes. Out of the 100 attempts, 98 (98%) were successful. Unsuccessful attempts (2%) were due to the lack of semantic meaning in the question (i.e., “If it is altered, forward physician p?”).

For the successfully annotated questions, between 1 and 3 Q-Codes were assigned. Nearly three-fifths (56%) of the questions were assigned 2 Q-Codes; 41% were assigned 1 Q-Code; and 3% were assigned 3 Q-Codes. There were seven instances where the question was assigned at least one Q-Code, while being unable to be coded with any ICPC codes.

A cumulative total of 159 Q-Codes were assigned to the 98 questions. More precisely, 97.5% of these Q-Codes were assigned at the sub-subcategory level, with 1.25% of the Q-Codes being assigned at both the subcategory and the sub-subcategory level.

The vast majority of Q-Code assignments were almost equally split between the QC (42%) and QD (37%) domains. Domain QT covered 21% of the assignments, while a single assignment was made to the domain QP. Four domains were not assigned: QE, QH, QR and QS. Six of the top 10 Q-Codes assigned belong to the Patient's Category domain (QC), 3 to the Doctor’s Issue domain (QD), and 1 to the Knowledge Management domain (QT).

5. Discussion

Of the top 10 assigned Q-Codes, only one category (QC22 "Women's health") overlaps partially with ICPC, and thus, does not provide much additional information about the semantic meaning of the question asked. Additionally, some questions represent a need for information on clinical topics not associated with a specific patient,
which could be used for future cases; "Continuous Medical Education" (QT23) was used for these questions. In fact, one of the reasons for the implementation of the telehealth system in Brazil is to provide medical education (Alkmim et al. 2012; Joshi et al. 2011). The remaining eight categories from Table 2 provide information that would be lost if a clinical KO system (i.e., ICPC) was used instead of the Q-Codes. Five of the remaining nine Q-Codes represent age groups (QC11, QC12, QC13, QC14, and QC15). Age of the patient is important, allowing the general practitioner to provide the best and most appropriate care to his/her patients. Three of the top 10 assigned Q-Codes deal with prevention (QD41, QD42, and QD43). This is not surprising, as health prevention and promotion is important to providing primary care in Brazil (Alkmim et al. 2012).

Table 1: Top 10 Q-Codes assigned

<table>
<thead>
<tr>
<th>Rank</th>
<th>Q-Code</th>
<th>Title</th>
<th>Number of assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>QC22</td>
<td>Women’s health</td>
<td>23</td>
</tr>
<tr>
<td>2</td>
<td>QT23</td>
<td>Continuous medical education</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>QD43</td>
<td>Tertiary prevention</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>QD42</td>
<td>Secondary prevention</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>QC14</td>
<td>Adult</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>QD41</td>
<td>Primary prevention</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>QC15</td>
<td>Elderly</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>QC13</td>
<td>Adolescent</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>QC12</td>
<td>Child</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>QC11</td>
<td>Infant</td>
<td>3</td>
</tr>
</tbody>
</table>

As a Work in Progress, only 100 questions have been manually annotated by the submission deadline.

One limitation can be seen in the use of manual annotation, which reduces the ability to index large data sets, quality and number of the resulting annotations; is time consuming; and requires number of actions, including inter-annotator agreement (Névéol, Islamaj Doğan, and Lu 2011). A second limitation is the present inability to publish the deidentified data set in a publicly available data repository.

6. Conclusion and Future Work

The use of Q-Codes to annotate Second Opinion Requests from rural Brazilian primary healthcare providers has been tested to show its feasibility in facilitating
communications and coding among rural healthcare providers when expressing non-clinical and contextual issues. Results show that Q-Codes add value capturing information that otherwise would be lost if using only clinical systems such as ICPC.

Future work would include the research goal is to annotate all 550 questions contained within the sample data set. Depending upon the availability of resources, it might be desirable to annotate the entire data set of 1,669 questions.

To overcome the limitation of manual annotation, future work could investigate the use of semi-automated annotation methods to assign Q-Codes to large data sets quickly and more efficiently. Other improvements can be the involvement of a second annotator to validate the dataset; and the use of DeCS to index the SOR to test its feasibility.

To overcome the lack of public access to the deidentified dataset, UFPE NUTES has a “plan to do this through a new component in our telehealth platform, but probably available by the end of 2018” (Magdala de Araújo Novaes, e-mail message to author, December 6, 2017).

Acknowledgments
The authors thank Nilma Andrade, Stefan Darmoni, Paula Rejane Beserra Diniz, Julien Grosjean, Ashwin Ittoo, Fernando José Ribeiro Sales, Robert Vander Stichele, and Marc Vanmeerbeek.

References


Luana Farias Sales, Luís Fernando Sayão

Use of conceptual relations for semantic integration of scientific publications and research data

Abstract
The context of this research is the 4th Scientific Paradigm, better known as e-Science. E-Science is characterized by the intensive use of computer networks, distributed digital repositories, and the reuse of research data. This study argues that a scientific publication can be enriched and be closer to the new forms of modern scientific knowledge generation if it is configured according to a model that, through semantic relations, links research data and datasets to conventional publication. Such a publication model that integrates scientific data and search information with results through semantic relations is called an enhanced publication. This paper will present a methodology used for the semantic integration of data and publications made available in a distributed environment. This approach is based on the use of conceptual relations to establish links between distributed digital objects. The results are a taxonomy of relations and an interface model to be computationally implemented.

1. Context
This research is developed within the context of an emerging scientific paradigm, known as e-Science or the 4th Scientific Paradigm. This new way of doing science is characterized by the intensive use of computer networks, distributed digital repositories, and the intensive generation and reuse of research data. The informational environment that has arisen as a consequence of these changes has had a significant impact on patterns of scientific communication, especially in terms of the means of communicating and disseminating research results. This work is the result of a doctoral thesis research project which is based on two premises: the first raises the need for a scientific publication model that can express and reflect the new pattern of generating scientific knowledge, rich in data, which can be integrated into publications; the second proposes that this can be accomplished according to the technological possibilities and patterns derived from theories of knowledge organization. These two premises embody the formulation of a hypothesis which argues that a scientific publication can be enriched and be closer to the new forms of modern scientific knowledge generation if it is configured according to a model that, through semantic relations, links research data and datasets to conventional publication. Such a publication model that integrates data and search information with results through semantic relations is called an enhanced publication.
2. Objective

The objective of this work is to present a methodology used for the semantic integration of data and publications made available in a distributed environment. Semantic integration is understood as the use of well-established meanings to forge relationships between digital objects distributed by different systems. Relationships or relations, as they are also called in the literature, can be broadly defined as "an association between two or more entities or between two or more classes of entities" (GREEN 2001, p.3). In the domain of knowledge organization, relationships are mechanisms used to indicate the existence of concepts that have some contextual similarity, thus providing more accurate retrieval. While such relationships allow users to become aware of the existence in the system of other documents that may be of interest to their purpose, they also allow for more efficient systems, ensuring consistency in "adoption of algorithms based on rules of association" (KIETZ et al. 2000).

3. Method

The methodology proposed for the semantic integration of data and publications consisted of the following steps:

1) Identification of the types of data that could be included in the enhanced publication and could be interconnected (e.g. measures, formulas, models, graphs, etc.). In this research, the study was specifically directed to the field of nuclear energy and the data developed by one of the institutes of the National Commission of Nuclear Energy of Brazil: the Institute of Nuclear Engineering (IEN).

The mapping of products generated by research activities was based on the completion of a questionnaire by leaders in the research areas. The result obtained is represented in the following table:
Table 1: Data generated by thematic areas of the IEN

<table>
<thead>
<tr>
<th>RESEARCH DATA</th>
<th>ENGINEERING AND TECHNOLOGY OF REACTORS</th>
<th>RADIOCHEMISTRY AND NUCLEAR CHEMISTRY</th>
<th>DEVELOPMENT OF TECHNOLOGY FOR COMPLEX SYSTEMS</th>
<th>DEVELOP. AND CHARAC. OF FUNCTIONAL AND STRUCTURAL MATERIALS</th>
<th>APPLICATIONS OF NUCLEAR TECHNIQUES IN INDUSTRY</th>
<th>NUCLEAR KNOWLEDGE MANAGEMENT</th>
<th>NUCLEAR INSTRUMENTATION AND CONTROL SYSTEMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measurement</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Survey Result</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Experiment Result</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Formula</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Equation</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Algorithm</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Image</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Video</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Audio</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Animated movie</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Film</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Photograph</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Database</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Simulation</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Nuclear Code</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Table</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Graphic</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Diagram</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>TEXT</td>
<td>3D model</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>---------------------------</td>
<td>----------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Reduced model</td>
<td>Y</td>
<td>N</td>
<td>S</td>
<td>S</td>
<td>N</td>
<td>N</td>
<td>S</td>
</tr>
<tr>
<td>Draw</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Metadata</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Questionnaire</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Interview</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Annotation</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Standard</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Model</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Certificate</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Laboratory notebook</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Transcription</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Correspondence</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Diary</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Field notebook</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>ARTEFACT</td>
<td>Specimen</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Sample</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Maquette</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>PROCESS</td>
<td>Standard operating procedure</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>workflow</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Protocol</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Test</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>OTHER</td>
<td>Phantom or mannequin</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
</tbody>
</table>

2) The second step was the identification of the types of publications that would have interlinked data (e.g. thesis, article, project, report, etc.). In this research, the publications were identified from the institutional repository of information.
existing in the IEN. The publication types are: papers, book chapters, dissertations of Master’s degree courses in IEN, learning objects, patents, reports, software, theses and dissertations produced by IEN employees, congress papers and presentations.

3) In a third step, other categories of information were identified that could be interconnected to form the enhanced publication (research data, e-print, project, person, organization). As in the first step, these categories were identified from completion of a questionnaire by the leaders in the research areas. The result obtained is represented in the following table:

Table 2: IEN e-prints by thematic area

<table>
<thead>
<tr>
<th>THEMATIC AREAS</th>
<th>ENGINEERING AND TECHNOLOGY OF REACTIONS</th>
<th>VIRTUAL REALITY</th>
<th>RADIOCHEMISTRY AND NUCLEAR CHEMISTRY</th>
<th>DEVELOPMENT OF TECHNOLOGY FOR COMPLEX SYSTEMS</th>
<th>DEVELOPMENT AND CHARACTERIZATION OF FUNCTIONAL AND STRUCTURAL MATERIALS</th>
<th>APPLICATIONS OF NUCLEAR TECHNIQUES IN INDUSTRY, HEALTH AND ENVIRONMENT</th>
<th>NUCLEAR KNOWLEDGE MANAGEMENT</th>
<th>NUCLEAR INSTRUMENTATION AND CONTROL SYSTEMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPRINTS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technical report</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Research report</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Application for funding</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Form</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Request for consent</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Blog</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Social network</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Research in progress Report</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Ethic committee</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Research project</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Evaluation of the funders</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Peer review</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
</tbody>
</table>

The above table shows the existence of other types of digital documents that also record the memory of scientific research and therefore deserve to be connected and offered to the user as a response to their subject consultation.

4) The fourth step was the definition of the roles played by each of these categories (e.g. for people: author, editor, advisor, etc.) The categories used were based on the EUROCRIS (2012) metadata scheme, as well as the conceptual definition of each of the entities identified and also of their roles.
EPRINT - digital format research objects used to formally and informally communicate results of academic research activities, such as articles, theses and dissertations. They include new documents and electronic media such as blogs and social media registrations.

RESEARCH DATA - research objects created experimentally, theoretically or by simulation. They include a broad set of objects generated by derivation, inference, analysis, annotation, and reformatting of other information objects.

PROJECT - research object consisting of documents that record technical, scientific, administrative, planning and budgetary parameters aimed at the development of scientific and academic research.

PERSON - actors involved in scientific research flows in related fields, researchers and professors, as well as in areas of management, support and funding of research activities, such as coordinators, directors.

ORGANIZATION - institutions or units of institutions, such as laboratories, departments and courses where academic activities are carried out or research, teaching, promotion or planning of scientific research.

5) The next step was to combine the identified categories by means of arrangement, forming potential paired relations, called categorical relations. The pairs found are shown in the following table:

Table 3: Categorial relations

<table>
<thead>
<tr>
<th>ENTITIES</th>
<th>EPRINT</th>
<th>DATA</th>
<th>PERSON</th>
<th>ORGANIZATION</th>
<th>PROJECT</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEMANTIC LINKS</td>
<td>EPRINT_EPRINT</td>
<td>DATA_DATA</td>
<td>PERSON_PERSON</td>
<td>ORG_ORG</td>
<td>PROJECT_PROJECT</td>
</tr>
<tr>
<td></td>
<td>EPRINT_DATA</td>
<td>DATA_DATA</td>
<td>PERSON_PERSON</td>
<td>ORG_EPRINT</td>
<td>PROJECT_EPRINT</td>
</tr>
<tr>
<td></td>
<td>EPRINT_PERSON</td>
<td>DATA_PERSON</td>
<td>PERSON_DATA</td>
<td>ORG_DATA</td>
<td>PROJECT_DATA</td>
</tr>
<tr>
<td></td>
<td>EPRINT_ORG</td>
<td>DATA_ORG</td>
<td>PERSON_ORG</td>
<td>ORG_PERSON</td>
<td>PROJECT_PERSON</td>
</tr>
<tr>
<td></td>
<td>EPRINT_PROJECT</td>
<td>DATA_PROJECT</td>
<td>PERSON_PROJECT</td>
<td>ORG_PROJECT</td>
<td>PROJECT_ORG</td>
</tr>
</tbody>
</table>

Table 4: Taxonomy of semantic links extract

<table>
<thead>
<tr>
<th>CLASSES OF RELATION</th>
<th>SEMANTIC LINKS</th>
<th>CLASSIFICATION</th>
<th>ONTOLOGY/SEMANTIC VOC.</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPRINT_EPRINT</td>
<td>Cites</td>
<td>Citation</td>
<td>CITO</td>
</tr>
<tr>
<td>EPRINT_DATA</td>
<td>Documents</td>
<td>Influence</td>
<td>CITO/Data Cite</td>
</tr>
<tr>
<td>EPRINT_PERSON</td>
<td>Is advised by</td>
<td>Property</td>
<td>VOC_IEN</td>
</tr>
<tr>
<td>EPRINT_PROJECT</td>
<td>Is result of</td>
<td>Meronym</td>
<td>DoCo</td>
</tr>
<tr>
<td>EPRINT_ORG</td>
<td>Is funded by</td>
<td>Property</td>
<td>VOC_IEN</td>
</tr>
<tr>
<td>DATA_DADO</td>
<td>Is version of</td>
<td>Derivation</td>
<td>Data Cite</td>
</tr>
<tr>
<td>DATA_EPRINT</td>
<td>Is figure of</td>
<td>Meronym</td>
<td>DoCo</td>
</tr>
</tbody>
</table>
This methodology was originally applied in the area of nuclear energy, but it may be applicable to other areas as well. The following exemplification considers a possible selection of resources, actors and relationships. There is an E-PRINT "A" authored by PERSON "B", who is the coordinator of PROJECT "C"; PERSON "B" works in
ORGANIZATION "D" as director; PROJECT "C" is financed by ORGANIZATION "E"; it is considered that E-PRINT "A" opens a discussion about data set "F" (DATA "F") that is presented in multimedia format; "A" is an item from the E-PRINT publication "G" and is commented on in the E-PRINT "H" blog of the author "B" and makes inferences about the E-PRINT "I"; PROJECT "C" generated the E-PRINT patent "J" which is described by "A". See the figure below which illustrates the example:

Figure 1: Model design for system implementation

The advantage of this type of integration is the possibility of at once recovering and integrally forming a distinct range of information about the same subject. In addition, formal relationships can be useful for knowledge processing. It is important to highlight that a semantic link always connects two entities that have a relevant role in the universe of scientific research, whether they are research results, people, institutions, projects or any other actor or research object identified as necessary for the representation of the domain where the scientific activities are performed. The relationships established between the various research objects and actors - that is, the research entities - can have their meanings attributed by different pre-existing ontologies.

4. Results

This is a theoretical study which resulted in a taxonomy of relations to be applied for integrating data, publications and information regarding the research on the creation of enhanced publications in the field of nuclear energy. Besides this taxonomy, the
work also resulted in an interface proposal using the taxonomy of relations to enable the integration of information in heterogeneous systems and consequently the integrated retrieval of information on a given topic. This interface can be seen in another work (SALES 2014). In the future, it is hoped to be able to implement the proposed interface computationally, thus integrating three systems: The Nuclear institutional repository, The Nuclear Current Research Information System - CRIS and The Open Journal Systems of Nuclear Energy.

Table 5: Relations Taxonomy for publication and data integration

<table>
<thead>
<tr>
<th>TYPE OF RELATION</th>
<th>SUBCLASSES AND EXAMPLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOGIC</td>
<td></td>
</tr>
</tbody>
</table>
| Relationship that shows genre-species hierarchies. | Is a  
|                  | Is a subclass of       |
|                  | Is a subgroup of       |
| MERONYMIC:       |                        |
| Relation between the parts of a publication | Is formed of  
|                  | Is part of adjacent   |
|                  | involves              |
| CITATION         | reveals the type of citation between documents |
| DERIVATION       | reveals the phase and stages of a document or time-based research process |
| PROPERTY         | reveals attributes, characteristics, role, or properties of a document |
| INFLUENCE        | reveals the impact, effect or action of one object or process on the other |

References


John Adetunji Adebisi, Babajide Samuel Afolabi, Bernard Ijesunor Akhigbe

A Value-based approach to modelling interoperability in Knowledge Organization Systems

Abstract
Existing KOS models are still evidence-based with a focus on how well they are holding up in terms of improvements. However, issues of correctness and reusability based on KOS conformance in terms of requirements re-engineering and metadata restructuring for interoperability gains towards better information access has not be given sufficient attention. This study, therefore, aims to provide insight into how to optimize interoperability in knowledge organization systems (KOSs) using an information system modelling approach. We have designed and implemented a unified value-based model using some identified mechanisms. Enterprise data information was integrated at the conceptual modelling level following the provisions in Diniz et al. (2006) to benefit KOSs in terms of interoperability with the goal of achieving information access. The identified components (or mechanism) required for the value-based model design were identified using appropriate fact-finding techniques, such as active stakeholder participation, interviews, record inspection, direct observation and searches from secondary sources. The design was implemented using extract-transform-load mechanism as a key technique, and the object-based and web-empowered methodology that uses Microsoft technologies such as SQL, ASP.net, web services and Transact SQL codes. Based on potential usage, the various modules that were integrated as a third-party value-based model were done by categorizing all extracted information to showcase seamless information access considering the level of interoperability achieved. This demonstrates the possibility of achieving seamless interoperability with legacy applications like KOS. In future, it will be interesting to implement a prototype of the proposed model, which is ongoing, in order to properly validate our claim.

1. Introduction
With today’s ever-increasing sources of online and offline data, interoperability is becoming more and more indispensable in order not to drown in data while starving for information (Ziegler 2004). In general, the goal of data integration is to combine data from different sources by applying a global data model. This often involves the detection and resolution of conflicts existing in the data schema to achieve homogeneity as well as provide a unified layer. The reason for data integration is two-fold: Firstly, given a set of existing data sources, an integrated view is created to facilitate data access and reuse through a single data access point. Secondly, given a certain information need, data from different complementing sources can be combined to provide a comprehensive basis to satisfy users information needs (Ziegler 2004). There is a remarkable history of research to draw from in terms of data integration for this paper to contribute further in the area of knowledge organisation system (KOS) interoperability. The spectrum of research ranges from early multi-database systems (e.g., Multibase) (Landers and Rosenberg, 1982), over mediator systems (e.g., Garlic) (Brown 2006) to ontology-based integration approaches (e.g., OBSERVER) (Knoblock 1996). What these approaches have in common is the autonomy of data sources that are
to be integrated. This was considered paramount. However, it came with an implication; users were not to be forced to adapt to any standard concerning the structure and semantics of the data they desire, something that was not even previously achievable.

From the literature, one can observe that; integration pertains to the exchange of meaningful information between systems regardless of whether the systems are designed to work together or not Mousa et al. (2014). While interoperability should have allowed the smooth sharing of information; the interpretation of incoming data and its presentation as received, and the preservation of its original context must also be possible (Bobby 2017). To achieve this, a number of techniques, algorithms and solution(s) have been developed (Lenzerini 2002; McCann et al. 2003 and Chohan et al. 2010). Moreover, data integration can be categorized into two types based on the nature of integration and the techniques. The first type is physical integration using data warehouse (DW) techniques (Kondabolu and Nasina, 2010). The second type is virtual integration that generally adopts data virtualization (DV) as its primary technique (Kondabolu and Nasina 2010; Eden 2007). For meaningful information exchange or integration to happen; providers and consumers need compatible semantics between source and target systems (Rosenthal, 2004). The aim often is to engender better ways of sharing data among systems especially within an enterprise. This means answering the question of what concepts would be used to achieve descriptive metadata for existing information; and, what definition(s) would be adopted to allow newly-built systems to interoperate. The goal will be to create a common platform that will ease the goal of data sharing for which both the new and old systems were built (Rosenthal 2004).

The concept of interoperability in enterprise applications (EAs) requires the proactive production of useful semantic agreement, and not just document correspondences among existing systems. This will help the enterprise satisfy the need for new requirements, such as the collection of new data or the establishment of new data sharing arrangements. This will also lead to the reduction of costs by cutting down unneeded semantics and diversity of representation (Rosenthal 2004). This should exclude concept definitions and relationships. For example, one may provide data on the aircraft that are operated by a country’s airline, while somebody also may want data on all the aircraft that regularly use the airports in that country. To properly interpret and integrate this type of data, one must understand what portion of information is needed to actually satisfy the foregoing information need. This study considers “interoperability” as it pertains to tolerating organizational “realities”, which is often ignored. This definition is needful to help narrow the practicality of this work to the central authority of the system to be integrated. As a result, we use the requirements of a third-party application to achieve the aim of our research. The objective is thus to design and implement a unified value-based model using some identified mechanisms that will serve as the components of the proposed model. Therefore, as in Diniz et al.
(2006), this paper provides insight on how to optimize interoperability in knowledge organization systems (KOSs) using an information system modelling approach. Third-party application interoperability requires specific activities, which are quite difficult to re-engineer. However, with the techniques employed in this paper, it was possible to simplify their metadata as well as bring more value to them. This has implications for legacy applications like KOS and, based on potential usage, it will be possible to make various modules in KOS interoperable, as with third-party applications. The paper is further structured as follows: section 2 is committed to reviewing the literature; 3 and 4 contain the theoretical foundation and the methodology adopted respectively (in the paper); and the proposed value-based approach to modelling data interoperability; finally, the paper’s conclusion is presented in section 5.

2. Literature

Some authors have attempted to apply theory of value to integration and interoperability. For example, the work of Keizer et al. (2015) used the domain modelling approach to resolve the problem of heterogeneity that often originates from differences in terminologies, domain coverage and root ontology. This challenge comes from the use of application profiling as a way to extend individual domain knowledge. This makes it difficult for a non-universal layer to be used as a common layer to address different application needs. Achimugu et al. (2010) proposed an enhanced model for the software engineering process of EAs. The goal was to contribute to how the critical appraisal of software architecture and methodology as a tool for efficient software engineering processes was carried out. The research provided vital insight as to how to remove the bottleneck that is usually faced during the integration of disparate systems for better interoperability. Considering the impact of software architecture with respect to system components, the study provided a guide to software architecture and a best practice example both of which were adequate to support our data integration model. This was important since, an in-depth look at the architecture of an EA is necessary to carry out a re-engineering exercise.

Both the work of Pokharel (2010) and Fatudimu et al. (2013) considered the place of a framework for the integrated mining of heterogeneous data in decision support systems. These studies highlighted the need to discover knowledge from both structured and unstructured data if completeness and comprehensiveness for data sharing is to be possible, while stressing the need to tailor interoperability towards knowledge sharing. Furthermore, Lindström and Polyakova (2010) considered a customer relationship management tool, which can be studied to make interoperability towards information access work. Based on the results from analysis of the tool, an integrated customer-centric model was developed. Thus, it is absolutely necessary to consider users’ needs to have easy access to information from the context of enterprise databases.
3. The theoretical foundation

The relational database model (RDM), which forms the theoretical basis of EAs, is limited in terms of data interoperability. When third-party applications that require re-use of data from other sources are involved, the RDM literally fails during operation (Lenzerini 2002). Software engineering researchers have proposed different approaches and software vendors themselves have intensified efforts at finding useful tools to help with the task of data integration. However, these efforts remain an art that relies on human labour, due to the absence of a comprehensive philosophical guide or appropriate theoretical foundation. This means that the overall modelling process of data integration still depends heavily on a data integrator's experiences (Brodie and Liu 2010). In other words, existing EA data models are designed to support only a single version of the “truth” within their domain of existence. Therefore, there will naturally be difficulties during data integration when the data collected from diverse sources are to be looked at from different perspectives or from different levels of abstraction (Liu 2012).

Shadow theory (ShT) serves as the theoretical foundation in this paper. The premise for the choice of ShT is because of its conception that whatever can be observed and stored into databases regarding a subject matter are just shadows, that is, mental entities existing only in the viewers’ cognitive structures (Liu 2012). These mental entities can be constrained by the viewers’ internal model of reality. This is particularly true of an implicitly or explicitly chosen perspective(s) of reality that can be formally represented. This option of being able to formally represent reality as postulated by ShT allowed the use of algebraic theory to model data operations based on their meanings and not just their logical structure. Point-free geometry aided representation of data within the semantic space. Since this can be decomposed or aggregated in different ways concurrently; W(hat)-tags were used to draw on shadows for their meanings; and E(quivalence)-tags were used to recognize which meanings can be treated as the same. Enterprise customer data integration was exemplified to illustrate the data model design and operational principles.

Ideally, data interoperability is contingent upon the issue of reliability and effectiveness criteria. To bring EA integration up to speed in this regard, a combination of ontological, methodological, epistemological and ethical means can be adopted to ensure best practice and see that the foregoing criteria are met. Based on software engineering practice like that highlighted in Lakatos (1978) and Gruner (2011), efforts have concentrated on applying ShT to the re-engineering of EA as an artefact. As a result, the methodological and epistemological side of software development has been pursued.

4. Methodology

The various components of the value-based model proposed in this work and used
to demonstrate the intended value-based approach to modelling interoperability in KOs were identified through fact-finding techniques. The techniques included active stakeholders’ participation, personal interviews, record inspection, direct observation and searches from secondary sources. The information necessary to implement each identified component were extracted using the extract-transform-load mechanism. This mechanism is a database re-engineering technique that work with webservices such as JSON-RPC, JSON-WSP, Web Services Description Language (WSDL), XML Interface for Network Services (XINS), Web Services Conversation Language (WSCL), Web Services Flow Language (WSFL), Simple Object Access protocol (SOAP) and Representational state transfer (REST). This study used REST and Active Server Methods (ASMX) of .Net for the information exchange layer, although WSDL and JSON were also supported as they may be required during a third-party application connection. Third-party application of EA modules was adopted; interoperability was achieved using third-party consumption and decision-making procedures. These were specified by categorising extracted information based on potential usage.

The research design used use-case of Unified Modelling Language (UML) as a design tool. The value-based model was implemented using the Rational Unified Process methodology. By this methodology the Microsoft SQL Management studio and ASP.NET web services were leveraged for implementation purposes. The data virtualisation layer resulting from the interoperability model was specified as comprising of modules that were further partitioned into sub-modules.

4.1. Value-based model

The architecture in Figure 1 exemplifies the value-based model. The architecture is translatable to benefit KOSs by optimizing interoperability. KOS components such as libraries can be organized as presented in the architecture in Figure 1. Keizer et al. (2015) also discussed a distributed database approach using the example as specified in the architecture; each third-party application, internal and external business application and others can be integrated into the enterprise application. This may require a new database designed for the purpose. Unfortunately, this can be complex and error prone to achieve using the conventional application today.

However, at the interoperability layer and the unified information exchange layer, web services are introduced seamlessly with little or no human interaction. The model combined existing knowledge with the knowledge gained from field work and case studies to provide stakeholders with a solution for effective decision making (in the form of associated business rules). For this study, the businesses are the protocols defined to ensure interoperability. For KOS, there will be a need for such protocols to fully allow the intended interoperability and they should be deployed at the value-based interoperability layer. The aim of the protocol is to ensure seamlessness Buyya, et al. (2009). Based on this, access to enterprise application data will be seamless since the
bottleneck of interoperability will have been removed.

Figure 1: A Value-based Integration Model for Enterprise Application showing the Interoperability layer deployed for third-party applications

4.2. Data interoperability

The proposed model was also developed following both the logical and physical foundation of a typical data warehouse. At this stage, the components that already exist in an organization (since we are considering EAs) were identified. The missing components were introduced. The relational and multidimensional nature of the databases used for data warehouses and their data marts was leveraged for ease of localization and interconnection of access tools. First, the logical and practical mode of architecture was defined as shown in Figure 1. This became the configuration for collecting the required data. In all, the system involved a central repository storing the data for an entire organization, an optional operational data store, data marts, and metadata repositories.

5. Value-based system architecture

The architecture presented in Keizer et al. (2015) continues to have a profound impact. It still serves as one of the de-facto standards for data interoperability. This study adopted and adapted it to show the novelty of the proposed value-based model. The value-based interoperability model, with its intended working parts as components, is presented as a systematic whole in Figure 1. The architecture shows how the model is deployed to work, especially in enterprise knowledge organisations. The model could also be deployed for business collaborations. This should be the same in all applications that require specific information for interoperability purposes. This is meant to ensure
that interoperability is managed at every point in time. As a result, at the integration layer, the model data will be the user data. The model provided proper infrastructure for data and application structure. It included the server, network, hardware and software components for connecting and communicating with third-party infrastructures. The technical architecture responded to the requirements of scalability, performance, availability, stability and security. The proposed model is thus robust, reliable, flexible, extensible and parallel.

6.0. Conclusion

We should highlight the fact that sections 3.0 and 4.0 were discussed in an instructive narrative and indeed the descriptive nature of this paper seeks to exemplify how KOS can benefit from interoperability when achieved in EAs. Our aim has been to hopefully whet the appetite of the knowledge organization research community concerning how to implement interoperability in KOS. This paper is therefore not prescriptive, which is one of its limitations. However, it proposes a value-based model capable of providing seamless integration. Clearly, since third-party applications require integration with tightly coupled EAs, they will seek to gain control of enterprise data and the schema structure will not be compromised. As an ongoing work, the value-based model is singled out as contributing to the development of a resourceful approach to knowledge sharing in the context of interoperability based on an integrated collaborative situation. In future, it will be necessary to implement a prototype of the proposed model. This we believe could lead to the value-based approach being considered capable of serving the functional requirements of legacy systems. Thus, the particular advantage of a value-based approach when compared to evidence-based methodology is the fact that it can be harnessed towards providing better interoperability for easy information access.

References


Erika Alves dos Santos, Marcos Luiz Mucheroni

VIAF and OpenCitations: cooperative work as a strategy for information organization in the linked data era

Abstract
The present reflection integrates doctoral research based on a bibliographical study. It suggests the interoperability between the cooperative management and information organization instruments VIAF/NISI and OpenCitations as a means of favoring standardization in the presentation of references and the use of SPAR ontologies as a way of introducing scientific references as facilitators for the broad retrieval of and access to information in the context of linked data.

1. Introduction
The organization of information has been influenced by the technological advances observed over recent decades. The volume of information has increased exponentially, enhanced by the democratization of technologies as well as the appearance of new storage media, including cloud media, thus facilitating the process of searching for and retrieving information, especially from the perspective of researchers. Moreover, this increase in both the supply and types of information sources demands a reaction from information science to make this volume of data searchable and recoverable. In this scenario, technology acts as the protagonist, simultaneously playing the role of villain and hero in the cause and effect relation regarding the volume of information produced versus the volume of information treated and recovered.

Against this background of transition from analog to digital, references have achieved prominence as emerging sources of information. However, in contrast to most other sources of information, references are usually written by researchers, who do not always master the techniques of documentary description. The wide range of existing bibliographic styles also contributes to the misinterpretation of such recommendations leading to potential erroneous registration of references. According to the CSL, there are more than 800 different bibliographic styles. Such variety can also generate difficulties in interpreting and using metadata recommendations presented in the form of references.

The philosophy of cooperative work, which has become an increasingly frequent practice in information science, is an alternative to help control how the presentation of references is standardized. The sharing of functionalities between VIAF and OpenCitations, which are both consolidated instruments of management and information control, are a potential alternative to enhance the use of references as a source of information, especially if applied under the use of ontologies, such as SPAR Ontologies.
2. VIAF: authority control in information retrieval

The Virtual International Authority File (VIAF) is a joint project of the Die Deutsche Bibliothek (DDB), the Library of Congress (LC) and the Online Computer Library Center (OCLC c2017), designed in Berlin in August 2003 during the IFLA World Library and Information Congress. Initially, the project aimed to establish a link between the authority records from the Library of Congress and the Die Deutsche Nationalbibliothek to form a virtual name authority system.

The longterm goal of the VIAF project was to link the authoritative names from many national libraries and other authoritative sources into a shared global authority service for persons, corporate bodies, conferences, places, etc. (Bennett, Hengel-Dittrich, O'Neil, and Tillett 2006, 4).

This objective was gradually achieved and currently the VIAF project has 52 contributor libraries distributed in 38 countries, including Brazil (Romanetto 2017). Today, the VIAF combines multiple name authority files into a single OCLC-hosted name authority service. The goal of the service is to lower the cost and increase the utility of library authority files by matching and linking widely-used authority files and making that information available on the Web (Bennett, Hengel-Dittrich, O'Neil, and Tillett 2006; OCLC c2017).

Standardization is one of the fundamental tools for the management and retrieval of information. Nevertheless, sociocultural aspects negatively influence this process: "it is common for different international cataloging agencies to establish the names differently or conversely, to use the same form of the name to represent different authors" (Bennett, Hengel-Dittrich, O'Neil, and Tillett 2006, 5). Variant forms of the same name constitute name matching problems in VIAF and consequently in searches, since the establishment of relations between the variant forms of authority names is fundamental to ensure that the search results are the same, no matter if the search is made from the authorized search term or from its variant forms. The establishment of the relations between equivalent search terms is an information science assignment and, therefore, transferring it to the researcher may represent a negligent act.

References are directly associated to citations, which give materiality and sustainability to scientific arguments. Citations, in turn, are directly associated to authority records, demonstrated, above all, by the fact that a characteristic shared by the more than 800 existing bibliographic styles is the use of the last surname of the author to indicate the main entry for a work, both in citations and references. This presupposes that VIAF attributes apply beyond the cooperative cataloging consortia and authority controls in the cataloging units, but also in the authority registries in references.
3. References as sources of information retrieval: the OpenCitations case

The scientific dissemination text is one of the most common ways of communicating research and may be presented in several forms: article, report, thesis, dissertation, book, among others. Because it is via text that the transmission of knowledge is proposed, the scientific text has characteristics that distinguish it from others, among which clarity, impersonality and cohesion stand out. The construction of scientific reflection is empirical and based on considerations previously consecrated by peers; that is, reflected in the scientific text by means of quotations. In addition, the scientific text should allow a reproduction of the method and the cognitive process carried out by the author of a scientific study. In this regard, references are protagonists and also constitute sources of accredited information regarding the subject addressed in the text. Thus, the careful writing of references not only represents the author's respect for his readers, but also attends to the requirements of a scientific publication.

Despite its importance, the elaboration of references is one of the scientific assignments in which researchers are less skilled. The complexity in which the normative documents that guide the writing of references are presented, added to the wide range of different bibliographic styles, means that incorrect specification of authority metadata (among other errors) in the composition of references has become the rule, not the exception.

The OpenCitations project, which was established in 2010 by the Joint Information System Committee (JISC), under the direction of David Shotton (Oxford University), was designed to change the face of scientific publishing. It aims to make bibliographic citation links as easy to use as Web links (JISC 2017).

The main work of OpenCitations is the creation and current expansion of the Open Citations Corpus (OCC), an open repository of scholarly citation data made available under a Creative Commons public domain dedication, which provides in RDF accurate citation information (bibliographic references) harvested from the scholarly literature. These are described using the SPAR Ontologies according to the OCC metadata model, and are made freely available so that others may freely build upon, enhance and reuse them for any purpose, without restriction under copyright or database law. (OpenCitations n.d.).

The OCC is explicitly based on SPAR Ontologies and other controlled vocabularies: FRBR- Aligned Bibliographic Ontology (FaBiO), Publishing Roles Ontology (PRO), Bibliographic Reference Ontology (BiRO), Citation Counting and Context Characterization Ontology (C4O), DataCite Ontology and Friend of a Friend (FOAF).

All the terms from the aforementioned ontologies are collected within a new ontology called OpenCitations Ontology (OCO). This is not yet another bibliographic ontology, but rather just a place where existing complementary ontological entities from several other ontologies are grouped together for the purpose of providing descriptive metadata for the OCC (OpenCitations n.d.).

OpenCitations represents an attempt to establish cooperative work in the registration
of knowledge in the form of references, which appears to be productive, considering previous successfully consolidated cases such as the cooperative cataloging programs. OpenCitations aims to provide precise citation data in RDF format, and to allow contributors to build, improve and freely reuse these data for any purpose, free from copyright or database rule restrictions. It also encourages the use of SPAR (Semantic Public and Referencing) ontologies and the treatment of authority metadata as expressed and manifested in standards such as in the FRBR. In addition, the use of SPAR Ontologies demonstrates that OpenCitations is aligned with the trend in applying Semantic Web and linked data resources in reference management, which is potentially beneficial both for information management and the interconnection of references and authority metadata, thus favoring scientific communication.

4. The use of SPAR Ontologies as a facilitating instrument for information retrieval

The Semantic Web, artificial intelligence and linked data are being consolidated as elements of automated reasoning as well as instruments of information management and organization. The Semantic Web allows data to be identified and interoperrated by computing agents and made comprehensible to people, promoting the discovery and use of decentralized information, in a contextualized way.

In this scenario, metadata are structured by data models, which allows the information spreaded in diverse and isolated databases to be interconnected and related, raising the range of applications and services.

[...]

The linked data (LD) is a Semantic Web application which allows the interconnection of data aiming to facilitate their discovering, usage and reusing, in a comprehensive way by machines and people (Serra, Silva, Santarem Segundo 2017, 386; Berners-Lee 2006).

Linked data are derived from the attribution of meaning, relations and interconnections to the data spread throughout the Web and result from the application of ontologies which, by means of concept representations and the establishment of relations among them, constitute

One of the ways of constructing an organized relationship among terms under a domain, favoring the possibility of contextualize data, making the process of data interpretation by the information retrieval instruments more efficient and facilitated.

Ontologies, as Semantic Web technologies, are “computational artifacts that describe a domain of the knowledge in a structured way through to: classes, properties, relations, restrictions, axioms and instances.” (Santarem Segundo, Coneglian, 2015, 227). The data enrichment through the application of ontologies favors the distinction of terms or names with different meanings, yet described identically, as well as grouping terms with common meanings described in different ways, even when stored in different locations. Thus, the semantic organization of the information dispersed in
the Web, based on ontological principles, potentializes search engine responsiveness
by means of the relations established between data spread in different sources on the
Web.

One of the few models developed originally for the description of bibliographic entities through
OWL 2 DL, the main characteristics of SPAR, one of the characteristics that distinguish it from
previous contributions, are firstly the creation of ontologies of sufficient expressivity to meet the
requirements of academic authors and publishers, and secondly the development of accompanying
presentation Technologies […] that enable the ontologies to be easily understood by potential users
such as academic researchers, publishers and librarians who, while expert in their own domains,
lack skills in ontology modelling and knowledge formalization (Peroni 2014, 151).

In the specific case of libraries, ontologies, through linked data, essentially contribute in two situations: first, the interconnection of the data stored in bibliographic catalogs raise the possibilities of relations or reuse of bibliographic metadata. The ontological treatment of bibliographic databases also makes this metadata researchable not only by the search engines in online catalogs, but also in the Web. Secondly, the divergences in authority registry by different institutions contradict the principle of interoperability and reduce the indices of information retrieval. However, the interconnection of this data increases the possibilities of its retrieval, first, by standardizing the ways of registering authorities, and, second, since the information treatment ruled by ontological principles assigns an interoperable and universal feature to searches. This exempts the researcher from previously determining the databases where searches should be carried out, considering that the relations between data extend both the scope of searches and, consequently, of results.

The possibility of reusing data present in ontologies leads to gains for users, as it favors the enrichment of the data present in the catalogs, and for librarians by allowing authors, subjects and other elements to be represented uniformly, with unique identification, and, thus, minimizing efforts for describing bibliographic metadata, in addition to increasing the quality of the data, which contributes to the possibilities of finding resources. (Serra, Silva, Santarem Segundo 2017, 387).

5 The cooperative work between VIAF and OpenCitations in the management of references

Several problems are reported in the literature (Davies 2012; Kratochvíl 2017) related to failures in the interpretation and use of standards destined for the elaboration of references. The unnecessary proliferation of standards in several institutions generates individual interpretations for each research unit which is reflected in the more than 8,000 bibliographic styles available at the Citation Style Language (CSL) database (CSL 2017), in contradiction to standardization principles.

Shotton (2013) points out that research practice suffers because access to citation data is currently so difficult, highlighting the need for these data to be open access and legally available for sharing in an academic citation open data repository; namely, OpenCitations. In this way, the purpose of OpenCitations may contribute to unifying
the ways of presenting references and its widespread use may, additionally, make the existence of such a vast number of different bibliographic styles unjustifiable.

Academic citation data includes the

“authority control that consists of an important tool of standardization and terminological control of great contribution to the quality of the functioning of the information search and retrieval systems in which the unification of the values established by VIAF consortium contributes to the aggregation of values covering linguistic and cultural variations, as well as in the sum of metadata elaborated by institutions that act in the national bibliographic control, what provides a major coverage over the publications in each country”. (Romanetto, Santos, Alves 2017, 586).

Beyond the wide range of distinct bibliographic styles and the disregard for the principles of authority control during the writing of references, there is neither any shared taxonomy that describes characteristics of other types of publications besides books and articles (Peroni, Dutton, Gray and Shotton 2015), nor identification of data about reference content, not to mention the lack of consensus among editors regarding the location of metadata registry in their publications. Moreover, the fact that there is growing use of non-traditional supports of information has altered the means of searching for and retrieving information, when tags, hashtags and DOIs have also been considered as search terms. However, such elements are disregarded by the most commonly used bibliographic styles and this potentially reduces the indices of responses to information searches. This situation poses a new challenge to information science besides developing and managing instruments of registry and management of human knowledge; namely, the integration and socialization of these tools in the scientific universe.

Irregular language, formats and registries may negatively impact the flow of scientific communication and demonstrate the importance of cooperative work even in the management and development of competence in reference elaboration, considering, above all, that although their writing demands specific knowledge in descriptive representation, references are commonly used in all knowledge areas. Since academic citation data includes authority control, the adoption of a cooperative work philosophy between VIAF/NISI and OpenCitations is a positively promising fusion.

Cooperative work has already been established in information science, more specifically in descriptive representation, as a facilitating instrument of information registry management, such as cooperative bibliographic catalogs. Therefore, the adoption of this method of work in reference elaboration does not represent an innovation, but an adoption of a successful model that shows signs of success in several aspects: a) in the uniformity of authority indication (persons, entities, events and titles) in references, bibliographic catalogs and also in electronic search instruments; b) in promoting the identification, localization and access to referenced documents; c) in the availability of a reliable consulting source for the extraction of standardized references;
d) in the rationalization of physical and human resources in the writing of references and control of bibliographic normalization; e) in the possibility of interconnecting bibliographic metadata by means of Linked Data and SparOntologies.

6 Final considerations

Metadata, the semantic web and authority file relationships are common concerns to VIAF, Open Citations, ontologies, references presentation, the FRBR format and also to cataloging. Even so, there is no evidence of cooperative work among such initiatives. Although the majority of papers are still published in the form of static PDF documents, it is necessary to create tools to facilitate the transition of the online paper from its present “horseless carriage” state into the scholarly communication equivalent of a Ferrari. (Peroni 2015, p. 267). Just like in VIAF and other cooperative work initiatives, the elaboration of references should adopt this working philosophy, considering the facilities provided by the technologies. The application of SPAR ontologies in the decoding of bibliographic data records, and the assigning context to references, potentially makes them as complete as cataloging records themselves if not more so. In addition, when it comes to efficiency of information retrieval, SPAR ontologies offer the possibility of metadata records readable by the RDF language and, therefore, compatible with the semantic web and with linked data technology. Solidity and integrity in descriptive and thematic metadata must be a concern not only in the cataloging field but, above all, in the presentation of references, considering they are the primary and accredited source of scientific information. SPAR ontologies may enhance the retrieval rate of documents selected for the purposes of thematic representation, raising references from the status of mere document lists into selected and interconnected information sources. This means inserting references into the scenario of “academic communication equivalent to a Ferrari”; that is, linked data.

References


Societal Challenges in Knowledge Organization
Author information for Knowledge Organization in the digital age

Abstract
Authorship is a critical topic in knowledge organization. In the Anglo-American cataloging tradition, authorship is often used as a characteristic for arrangement and as an access point for retrieval. There is a body of literature in knowledge organization that examines the meaning of authorship. Through literature review, we can see that authorship is a cultural, social, and temporal concept, and author information in the cataloging tradition does not present all the author functions. This paper discusses challenges concerning the collection, recording, and representation of author information in the digital age, as well as the consequences of enriching author information. We argue that more information is not necessarily better for cataloging in the digital age.

Introduction
Authorship is a critical topic in knowledge organization. In the Anglo-American cataloging tradition, authorship is often used as a characteristic for arrangement and as an access point for retrieval. Author information serves both the collocating and identifying objectives of library catalogs (Cutter 1876). Tracing back to Roland Barthes’s The Death of the Author and Michel Foucault’s response What is an author? previous studies review the authorship discussion and examine authorship in different times, social and cultural contexts, and across various forms of works.

Foucault points out that the concept of author goes beyond the attributes of an individual. Identifying someone as the author of a work tells us more than the person being the content creator. Author also indicates what Foucault calls the author function, which is a complex mass of characteristics influenced by social, cultural, and temporal factors (Foucault 1970). An author may indicate certain value, a style, a mode of discourse, or historical events. Using Freud as an example, Foucault shows how an author, as a founder of a discourse, may have influence on other authors and works that go beyond the individual’s intention. The author function may group, differentiate, and relate texts, as well as characterize a discourse or specify ownership (Martínez-Ávila et al. 2015; Foucault 1970). Expanding on this, thinkers in the knowledge organization context explore and identify definitions, characteristics, and functions of the author. Through examples of studies of superworks like the Seven Epitomes, Abelard’s Works, and The French Chef (Smiraglia et al., 2011; Smiraglia and Lee 2012; Smiraglia et al. 2013; Martinez-Ávila et al. 2015; Lee 2016); as well as studies of contemporary projects like the Europeana, AustLit, the American Civil War: Letters and Diaries, and DBpedia (Moulaison et al. 2013; 2014), we can recognize some prominent observations. First, authorship is not a mere attribution. It is a culturally, socially, and temporally bound concept which is related to information objects (e.g., books). We can
explain this with an example. Recording William Shakespeare as the author of Romeo and Juliet addresses more than identifying him as the person responsible for the content and creation of the work. Our understanding of Shakespeare and his time provides hints for interpretations of his works, and it may shed light on relationships between Romeo and Juliet and other works or people. The second observation drawn from the knowledge organization literature is that author information, in the cataloging tradition, does not present all the Foucauldian author functions. The cataloging tradition treats author quite simply, as a text string. For the purpose of authority control, catalogers often use the birth date and death date, or the occupation, as qualifiers to distinguish different identities with identical names. Even if we check the name authority record, the available author information and author functions are still limited. While the classificatory function and ownership aspect of Foucauldian author function is present in this context, we do not have access to style, value, historical background, and other aspects important to Foucault (Foucault 1970).

The U.S. cataloging standard RDA (Resource Description and Access) places more emphasis on relationships. And because of this, catalogers are recording more author information. However, the author function introduced by Foucault (1970) identifies particular relationships and expansive influences of an author which go beyond the coverage of current cataloging practice. In the works reviewed above, scholars acknowledge the richness of authorship and recognize the limited recording of author information in cataloging practice. For example, Moulaison et al. (2013; 2014) elaborate on how events may be a critical aspect of an author’s life. An author might meet people at events and change their relationships, thoughts, or works. However, few knowledge organization systems record and present event information as part of the author information. The authors use NNBD Mapper to provide an example of visualization of event information of a person. They suggest recording more author information and making them searchable for users to realize Foucault’s author function in current knowledge organization systems.

Based on the literature reviewed, this paper recognizes the motivation of pursuing better presentation of author function in knowledge organization systems, but argues that more information is not necessarily better for cataloging in the digital age. To do this, we discuss three challenges concerning collecting and recording author information in the digital age. The first challenge focuses on representing author information. The second challenge focuses on collecting author information. The third challenge is concerned with the consequences of enriching author information.

**Representing author information**

The first challenge focuses on the representation of author information. How can we address the cultural and temporal nuances of authorship in a “standardized” manner, so institutions can share metadata? On one hand, we recognize the meaning of authorship
differs by time, culture, and social context. On the other hand, we need a modicum of standardization to enable metadata sharing.

Lee (2016) identifies six types of authorship presented in the *Seven Epitomes* (*Qilue*), which is, for imperial China, “the first classified catalog and the one that established the model for bibliography”. These six types, namely composer, speaker, transmitter, exegete/commentator, patron, and editor reflect the definitions of authorship at the time of the catalog (i.e., the Former Han dynasty, 206 B.C.E. - 8 C.E.), in the Chinese culture, and for particular social contexts, like the catalog for the imperial library. While some types of authorship, such as *composer*, may be covered by *author* in RDA, others may map with different roles contributed to the creation of the work. For example, we may map *patron* in the *Seven Epitomes* to *sponsor* in RDA. Further, there is no guarantee of a one-to-one mapping. In Lee (2017), the author discusses six mapping issues between Chinese and English role designators. There are Chinese terms indicating mixed English roles (one-to-many mapping), old or ambiguous terms, synonyms (many-to-one mapping), homographs (one-to-many mapping), homographs across languages, and Chinese terms with no match in English (one-to-zero mapping). Some of the issues are complicated and can be obscured by translation. The combination of cultural, temporal, and social nuances may be lost or misunderstood in the activities of translation, mapping, and the cataloging process.

The two examples above only present part of the challenges of representing author information. Other issues may arise when catalogers use controlled vocabularies to represent author-provided information. Thompson (2016) conducts content analysis of name authority records of authors who self-identify as trans. She shows how the limitations and values embedded in controlled vocabularies and cataloging rules may be imposed to author information. One of her examples is assigning start and end dates to a gender identity. This approach assumes that an individual’s gender shifts from one to another, rather than being fluid. The assumption imposed by the descriptors may not faithfully represent an author’s self-identification.

To improve the presentation of the author function in information systems, we may want to go beyond mapping types of authorship and role designators between different times, cultures, and societies; or go beyond representing author information using a limited controlled vocabulary. We would expect systems to provide richer author information, such as the intellectual tradition of an author, an author’s relationships with other authors, or the influences of an historical event on an author. The challenge of representing author information is one of improving the flexibility and hospitality of author descriptions to record context-specific expressions, while remaining interoperable across information systems.
Collecting author information

The second challenge focuses on collecting author information. When it becomes easier to collect information, make linkages, and specify relationships in the digital age, what are the concerns of collecting and providing author information? Moulaison et al. (2014) mention ethical concerns of both recording particular author attributes and the challenge of making the author information searchable. Thompson (2016) addresses privacy and safety concerns about collecting and providing certain author information, such as gender identity.

In addition to these concerns, another issue lies in considering the appropriate sources of information. What author information would be both practical and ethical in scope for catalogers to collect? One example case is Wikipedia, which relies on Wikipedia editors to populate information from a variety of information sources. According to the Wikipedia: Identifying reliable sources page, “Wikipedia articles should be based on reliable, published sources.” In addition, “contentious material about living persons (or, in some cases, recently deceased) that is unsourced or poorly sourced—whether the material is negative, positive, neutral, or just questionable – should be removed immediately and without waiting for discussion” (Wikipedia 2017).

We also have a case like the Library of Congress Demographic Group Terms (LCDGT), which includes eleven categories of terms (i.e., age; educational level; ethnic/cultural; gender; language; medical, psychological, and disability; national/regional; occupation/field of activity; religion; sexual orientation; social) to describe the demography of creators, contributors, and intended audiences of resources (Library of Congress 2017). Catalogers can assign LCDGT in both bibliographic records and authority records. The purpose of LCDGT is to cover the demographic terms in the LCSH. With the development of another thesaurus – the Library of Congress Genre Form Terms (LCGFT), the goal is to separate both demographic terms (LCDGT) and descriptors of what a resource is (LCGFT) from subject headings which describe what a resource is about (LCSH). According to the manual of the LCDGT (Library of Congress 2017), the practical and ethical guidance given is that the source of information for LCDGT should be limited to authors’ self-provided information in the piece described. The manual instructs catalogers to use other information sources, including previous works by the same author, with caution. It acknowledges the complexity of author information and how identity may change over time. However, the guidelines assume that authors provide information in the resources described. It is an assumption we cannot take for granted. In order to publish a work, there may be censorship, either imposed or self-applied. Further, there are questions concerning the maintenance of LCDGT. Following the guidelines, the LCDGT, which describe the author in a bibliographic record, function like a snapshot of the author’s self-identified demography at the time of publication. However, in authority records, catalogers record
author information over time. How practical it is to update the LCDGT in authority records if demographic information changes? Could what is unveiled be veiled again?

In the digital age, there is a tension between the ethics of respecting authors’ privacy and providing richer author information. Both aforementioned cases address the ethical concerns of collecting personal information with caution. The LCDGT case limits the information source to author-provided information. The Wikipedia case allows a wider variety of published information sources; and it has guidelines and policies on reliability and verifiability of information sources. However, the evolving approaches for publishing information may create grey areas. Guidelines which limit the sources of information to published works or publicly available information may not be as clear-cut as we expect.

**Consequences of enriching author information**

The third challenge is concerned with the consequences of enriching author information. Does the amount of author information available influence the visibility and access of the authors’ works? By populating author information and making connections between authors and works, there is good intention to improve the selection and interpretation of works. For instance, information about an author’s relationship with a particular school of thought could help readers interpret the author’s works. As Moulaison et al. (2013) argue, knowing the author’s age at the time of publication may influence the selection of works.

Besides the improvement of selection and interpretation of works, in the digital age, richer author information may result in more access points and links to an author’s information, and thus increase visibility of an author and an author’s works. However, if richer author information leads to higher visibility of an author’s works, it may render other authors less visible. And this can be seen as an ethical issue. We may want to identify the factors influencing the recording and searchability of author information. For instance, does the language and script matter? If author information in some languages were more visible than others, we could examine the knowledge organization systems to detect influential factors other than reader preferences. Is a system hospitable for information in different languages? Would international collaborations complement the local system? An expanded version of VIAF (the Virtual International Authority File) might be one possibility (Online Computer Library Center 2018).

One other concern is how the presentation of rich author information influences contemporary authors, and what are the ramifications of this to non-contemporary authors. Generally, it is easier to enrich author information for people who are currently active than for people who wrote centuries ago. If the recording of author information systematically emphasizes some groups (e.g., contemporary authors) and render the other groups (e.g., non-contemporary authors) less visible, enriching author
information may lead to another ethical issue.

In the digital age, we have better abilities to enrich author information. We have access to more detailed information about some authors. Technologies enable us to add more links between authors and works across languages, regions, and times. It becomes easier to populate information and create links with efficiency and at a large scale. However, we should examine the approaches and consequences of enriching author information, and take different stakeholders into account. Some stakeholders include authors of different time periods, regions, cultures, and languages; and information professionals who aim to describe authors in both ethical and practical approaches; as well as people who find, select, and interpret works using author information. Since not all stakeholders are equally vocal, it is important for information professionals to identify them and their stances, and think through the consequences of enriching author information.

Conclusion

The purpose of raising these challenges is not to deter the pursuit of presenting the author function in knowledge organization systems. Instead, the goal is to increase awareness of some concerns, and encourage discussion and studies for approaches to ameliorate potential issues. In the digital age, with the ability to share, collect, and link information at a large scale, it is important to ask if providing more information is necessarily better.

References


Thompson, Kelly J. (2016). More than a Name: a Content Analysis of Name Authority Records for Authors Who Self-Identify as Trans. Library Resources & Technical Services, 60(3): 140-155. Available at: https://doi.org/10.5860/lrts.60n3.140.


Challenges of organization and retrieval of photographs on social networks on the Internet

Abstract
The purpose of this study was to analyze the evolution of the behavior of social network users on the Internet regarding the storage and retrieval of photographs, and how knowledge organization has been introduced in a natural and progressive way, through platforms and applications, in their day-to-day lives. This makes the experience of photograph sharing quicker, simpler and more attractive. Through the observation of how the sharing, storage and organization of photographs in these environments is done, it is our aim to record and contribute to a study of the impact on personal and collective memory in a society that has been using social networks as a place to store personal photographs and records of daily activities and events.

Introduction
It has been only 27 years since Tim Berners-Lee made the World Wide Web available to the public. Since then, the Internet has become popular and today is an integral part of the daily lives of half the world’s population, totaling 3.7 billion users (Hootsuite 2017). As Internet connection increased in reach, quality and speed, users and developers began exploring more functions and Internet applications, which caused the emergence of many sites with high user interactivity, among them, social networks. Since 2004 this scenario has been known as Web 2.0, characterized by mechanisms of collaboration directed to the user, where usability predominates, such as the use of free language, folksonomy, and keywords in tags. Despite synonymy and polysemy problems, it has high usability and the more it is used, the more useful it becomes (Moreiro Gonzalez 2011, 35). Globally, since 2007 the time spent on social networks has surpassed e-mail time. Similarly, the number of social network users has overtaken those of e-mail since 2009. Thus, the most important activity on the Internet occurs through social networks, and these sites are becoming platforms for all types of activities besides friendships: e-commerce, education, marketing, media, entertainment, sociopolitical activism, among others (Castells 2015, 40).

Digital network technologies allow people and organizations to manage their own content and messages and share them with friends, family, customers, and so on. A significant part of this is done through photographic images, which are used for various purposes, such as informing, communicating, remembering and, by not requiring words or a language to be understood, they have almost universal reach. According to Layne (1994, 583-588), an image can say something that a text alone cannot, and perhaps for its ability to communicate so much in such a short time, photography has been used so much and in so many ways on the Internet.
With the increase in the available space on Internet servers and the increasing amount and speed of information to which we are exposed daily, users and developers have begun to look for means of administering the content that traffics on the web. Parallel to this, the growth of personal photographic files on social networks on the Internet requires urgent attention and intervention as to the security and integrity of these documents and to the consequences in individual and collective memory. In this context, knowledge organization brings concepts that can solve one of the biggest challenges of the Internet, the organization and retrieval of photographs on social networks. Knowledge organization and representation schemes can help users and developers to save time and space on the web, optimizing the search, retrieval of information and access.

The purpose of this study is to analyze the behavior of social network users on the Internet in relation to the storage and retrieval of photographs, and how knowledge organization has been introduced in a natural and progressive way, through platforms and applications, in their day-to-day lives, thus making the experience of photograph sharing quicker, simpler and more attractive.

Accordingly, we performed a historical and analytical study of the use, sharing and custody of photographs in the main social networks created particularly for this purpose, a bibliographical review of how some knowledge organization instruments have been used in social networks on the Internet for organization and retrieval of photographs, and to understand and analyze social network users’ behavior, we conducted an online questionnaire about their experience regarding the use of photographs on the Internet. The questionnaire was available online between 13th and 29th December 2017 and obtained 2,126 answers. The Google Forms App was used and the dissemination was done mainly by social networks Facebook and WhatsApp. Considering that the number of social media users in the world is currently 2.8 billion (Hootsuite, 2017), our research presents a statistical sampling error of 2% with a 90% confidence level.

Small revolutions of the main social networks

Facebook is characterized as a social network facilitating the exchange of experiences, events and ideas, where photograph sharing has increasingly expanded both in terms of quantity and significance. The history and origin of Facebook is intrinsically linked to the sharing of photographs on the Internet. Its precursor, the Facemash, created in 2003 by Harvard students, consisted of a website to elect the most attractive student by presenting side-by-side photographs. Initially exclusive to college students, only in 2006 was Facebook released to anyone over the age of 13 with a valid email address. By 2012, the social network was sharing 300 million photographs per day (Kpcb 2013) and it is currently the largest social network on the planet, having reached the 2 billion user mark at the end of June 2017. (Forbes 2017).
In 2010, Brazilian Mike Krieger and Kevin Systrom created Instagram. Initially, it was only available for the IOS system and the practicality of applying filters in the photographs was so successful with iPhone users that, at the time, it compensated for the poor camera. The photographs captured through Instagram were done in square format, in reference to the elitist Polaroid. In addition, it was possible to share the images directly on Twitter and Facebook. It took only 3 months to reach the milestone of 1 million users and 1 year to have 15 million subscribed users. In 2012, the mobile app became available on the Android system and reached 27 million users worldwide (Correia and Moreira 2014). On the same occasion, Facebook founder Mark Zuckerberg disbursed $1 billion to purchase the Instagram brand and declared “For years, we've focused on building the best experience for sharing photos with your friends and family” (The Guardian 2012). In 2017, Instagram reached 800 million users.

On Instagram, hashtag was consecrated. Originated on Twitter, it consists of an indexing mechanism that allows the user to create a hyperlink with similar content.

In mid-2008, Twitter, one of the main social media networks, established the indexing system called 'trending topics', a tool that allows the grouping of posts by topic, articulating certain words, phrases or expressions preceded by the hashtag symbol "#". (Moura and Mandaj 2014, 6, our translation).

More than a billion people in more than 180 countries use the WhatsApp messenger to keep in touch with family and friends through their smartphones. In addition to being a social network, this mobile app released in 2009 changed the way people use the cellphone to communicate. People started calling less and less and sending messages more and more. Not only because it is cheaper but also because of the features that improve communication, such as sending photographs, movies, text documents in various extensions, location on the map, links and so on. And with the advantage of being able to do this in groups of up to 256 participants. Due to the high cost of mobile phone services at the time of its launch, adherence to WhatsApp was so great, that telephone companies had to re-adjust their billing plans to remain competitive in the market.

Another social network that brought disruptive innovation to the virtual relationship was Snapchat. In this mobile app, users, equipped with interactive editing tools, send photos and videos to a friend and program the time the images will be visible – at most 10 seconds. After this time, the images are automatically deleted and cannot be viewed again. If a follower takes a screenshot, to copy the image, the user who sent the image is notified. This volatility in information is a paradigm break because it removes the memory recording function of photography, which in this context is only a communication tool. This mobile app was very successful with the youngsters due to the possibility of exposing the content to specially selected users for each snap and without leaving a trace. The success of Snapchat was so great that Facebook, inspired
by this way of communication, created *stories*, a profile status where users can upload photos or videos that are available for viewing by friends and followers for up to 24 hours. Facebook also made *stories* available to its other social networks, Instagram and WhatsApp, exponentially increasing the amount of photographs that are intentionally deleted every day.

Flickr is a web site hosting and sharing images like photos, drawings and illustrations. Launched in 2004, it innovated in the treatment of digital photographs by allowing new ways of organizing photographs and videos. Characterized as a social network for its high level of interactivity with users, it adopts the popular categorization system of files through *tags*, which are considered a form of metadata, attributed to the respective photographs by the users themselves. Thus, the search of images becomes an easy and agile process. The site offers a list of tags to be assigned, providing an indexing experience for the users. In addition to allowing the user to create their own tags to index photos and thus benefit from folksonomy, defined by Moreiro González (2011, 46) as a cooperative indexing that adopts keywords to represent information spontaneously and freely, Flickr also allows users to organize their own photos through albums and group them into collections.

Released in 2010, Pinterest is a social network of image sharing that works as a framework of inspirations and ideas for a multitude of subjects and categories. The proposal is that the platform resembles a frame, where they fix the things that the user likes the most, known as *pins*, that the user can group into collections known as *pinboards*. Pins are classified into very varied pre-established thematic categories: recipes, crafts, travel, photography, pets, marriage, etc. These categories help in the search and retrieval of content, allowing users to organize the inspirations they find on the Internet to be accessed in the future, as well as share with other users. In addition, the user can follow specific subjects shared by other users, without necessarily seeing all the shared content, as in other social networks. This demonstrates that knowledge organization can contribute to the organization, diffusion, search and access of photographs in social networks.

**Indexing photos on social networks**

According to Slype (apud Araújo Jr. 2007, 21, our translation), indexing “is the activity that consists of representing the content of a document or an analytical query, that is, enumerating concepts and / or words". Araújo Jr adds that the conversion of the language of the document into a documentary language will involve the analysis of the document by the indexer, in order to identify and select keywords that may represent the content, so that the document is later retrieved. (Araujo Jr. 2007, 21).

When this involves photographic documents, there is a need to transpose the visual into the verbal. Smit (1996, 29) states that the representation of a photographic image
cannot be a simple automatic transposition of documentary analysis procedures developed for the text, because the statute of the photographic image is different from the text and because the use of the photographic image is not only due to its informational content, but also because of its photographic expression.

Among the different approaches to knowledge organization, the authors Brandt and Medeiros (2010, 114) believe that folksonomy can be analyzed as a scheme of knowledge representation centered on the user because it has two of the main characteristics of this approach. The tagging process that generates the system terms is totally carried out by the user and another essential aspect is that the term is chosen by the user.

Although folksonomy falls short of the principles of knowledge organization due to polysemy, lack of controlled vocabulary and other factors, we cannot rule out its importance in the process of representing information. Analyzing theoreticians like De Mai, Hjorland, Foskett, among others, the authors (Brandt and Medeiros) believe that no other form of knowledge representation is as directly linked to the cognitive world and to the needs of the user as Folksonomy. On this, Ferreira points out that:

People who use the networks to publish their photographic content feel, while the information professionals seek, objectively, to date, allocate and narrate the events. One is not more important than another [...] both are important and may also be allies in the process of representing information (2017, 11, our translation).

Considering the polysemy of meanings that a photograph can assume, the hashtag, allied with folksonomy, assists in the understanding of photography from the point of view of the producer, and whoever shares the image. Ferreira (2017, p. 2) defines the photographic image as a composite language that can trigger several objective and subjective readings and interpretations. In fact, photography, when presented without a caption, can be interpreted in many ways by different individuals, or even by the same individual at different times of his life. However, the photograph, tagged by the author himself, carries a production context that guides the viewer’s interpretation to the message that the author wants to convey.

**Analysis and results**

The questionnaire showed that the same promising scenario for photograph diffusion and retrieval can be very dangerous when it comes to memory preservation. Out of 2,126 respondents 1,336 (62.9%) reported having lost photographic records due to social network extinction (Fotologs, Orkut, and so on). Despite this, they continue to trust the custody of their photographs to the social networks on the Internet. 50.8% of users, 1,081 respondents, said they have photographs saved only in social networks, without backup in their own support. In addition, the number of photographs produced with the intention to communicate rather than to save, grows every day. The feature of
photography as a communication resource has recently gained prominence with Snapchat and all indications are that it will be increasingly exploited, as several social networks have adopted the stories, inspired by Snapchat. The questionnaire revealed that 33.7% of social network users use Snapchat and 71.9% share photographs in stories. However, more than 40% stated that they do not save this content that will be automatically deleted in 24 hours, corroborating the idea of using photography as a communication tool without the intention of recording memory.

In this context, more care must be taken in the treatment of photographs that one wishes to preserve, so that they do not get lost among the others. With the constant increase of connection speed and space availability on the Internet, the amount of photos uploaded and shared has increased exponentially. Thus, it becomes increasingly difficult to find a photograph without the aid of any tools. 51.3% of the respondents stated they already had difficulties in recovering a photograph and they gave up looking for the picture before finding it.

Indexing is an important photograph retrieval tool for social network users and, in a way, assists in the personal and collective memory maintenance until more accurate technologies for retrieving image content are developed. Until then, it would be interesting to guide and educate users on how indexing can be a tool in the present, assisting in communication, representation and organization and, in the future, in the recovery and access to photographs and memories.

70% of the questionnaire respondents reported using hashtags for various purposes, and 66% of them use it when sharing photographs. Among the purposes of use, 18.2% use it with the intention of creating an access medium to the photograph through a search, 18.4% use it with the intention of linking the photograph to other similar ones, 30% use it with the intention of describing the photograph, and 25% as a language resource; a way to express themselves. The low percentage of hashtag usage as a photo recovery tool is confirmed by almost 35% of respondents who declared they do not consider it advantageous to use hashtags in indexing photographs and the 13% who declared not knowing how to use the hashtags.

These results demonstrate the need to clarify to social network users the advantages of conscious indexing for photo retrieval. An interesting fact is that the use of hashtags on Instagram (71%) surpassed the use of hashtags on Twitter (23%), the network responsible for inserting indexing through the hashtag in social networks. This shows there is more interest by social network users in giving some kind of treatment to their photographs than indexing other content on the Internet.
Conclusion

Due to the large amount of photographs on social networks on the Internet, and the increasing number of photographs that are intentionally lost every day through stories and apps like Snapchat, we observed a change in the way photographs are used on the Internet. Increasingly used for communication and less as a record of the present to be revisited in the future.

We observed that social networks are introducing knowledge organization instruments into the daily lives of users in a natural and progressive way to meet the demand for photograph recovery and, consequently, demonstrating to society the value of knowledge organization for information management and preservation of social memory.

Folksonomy was decisive in this process because, by allowing the user to freely classify information, it ensures the best description of the production context and that the content will naturally be related to words that users believe to be linked to them, that is, the words that are most likely to be used when searching for certain content. In this way the platform moderator is freed from having to master or have in-depth knowledge of all subjects.

References


*Instagram tem 800 milhões de usuários ativos por mês e 500 milhões por dia* (2017). Available at: https://g1.globo.com/tecnologia/noticia/instagram-tem-800-milhoes-de-usuarios-ativos-por-mes-e-500-milhoes-por-dia.ghtml.


Challenges to Knowledge Organization in the era of social media: the case of social controversies

Abstract
In this paper, we look at how social media, in particular Twitter, are used to trigger, propagate and regulate opinions, and social controversies. Social media platforms are displacing the mainstream media and traditional sources of knowledge by facilitating the propagation of ideologies and causes championed by different groups of people. This results in pressures being brought to bear on institutions in the real world which are forced to make hasty decisions based on social media campaigns. The new forms of activism and the public arena enabled by social media platforms have also facilitated the propagation of so-called “post-truth” and “alternative facts” that obfuscate the traditional processes of knowledge elaboration which took decades to arrive at. This poses serious challenges for knowledge organization systems (KOs) that the KO community needs to find ways to address.

1. Introduction
Controversy arises as a result of the expression of a disagreement between different rationalizations or between different conceptions of the social world in which it unfolds (Romain Badouard et al. 2013). Controversy appears as a counter-speech, that is, it cannot happen without the pre-existence of a first speech that it will counter by antagonistic discourse (Ruth Amossy et al. 2011). The major issue of social controversy is to change its trajectory, to "move the lines" and to question self-will and beliefs in societies (Romain Badouard et al. 2013). To do this, protagonists of controversies exchange arguments in which they defend their position or thwart the opposing position (Patrick Charaudeau 2015). As protagonists of a controversy who are not ordinarily in the public eye have difficulty in accessing mainstream media (digital or print press, television) in order to air their opinions, they increasingly turn to “alternative media”. Social media platforms have become the new public arena that enable and empower ordinary people in disseminating their opinions, beliefs and views in a free manner without the usual checks and verifications that mainstream media and knowledge organization systems demand. The mainstream media are now playing “catch-up” with social media and have been relegated to being their echo chamber in relaying news first divulged on social media. Because of the rapidity of propagation of arguments and opinions put forward on social media, it is very difficult to counter false or approximate knowledge with the usual processes of knowledge verification and rebuttal. As a result, the notions of information and verified knowledge have never been more challenged than in the social media era where people’s beliefs are fashioned more after their network of friends and followers than on scientifically established knowledge laid
down in *bona fide* sources such as books, encyclopedia, scientific literature or expert opinions. Such is now the impact of unverified assertions and “alternative facts” in real life affairs that the Oxford Dictionaries chose “post-truth” as its word of the year in 2016\(^1\). Post-truth is defined as “a political culture in which debate is framed largely by appeals to emotion disconnected from the details of policy, and by the repeated assertion of talking points to which factual rebuttals are ignored”\(^2\).

In this paper, we look at how social media, in particular Twitter, was used to trigger, propagate and regulate social controversies in a way that challenged real-world knowledge about the issue being debated. Twitter has become a useful channel for the ‘social infomediation of news’ (Smyrnaios and Rieder 2013). As a case study, we look at the controversy resulting from a speech given by the beleaguered Nobel Prize winner, Sir Tim Hunt on the 8\(^{th}\) of June 2015 after a luncheon with women scientists and journalists in South Korea\(^3\). The controversy was triggered by a tweet sent by Connie St Louis who was then an adjunct professor of scientific journalism at City University London and was attending the event:

« *Nobel scientist Tim Hunt FRS* @royalsociety says at Korean women lunch « I’m a chauvinist and keep ‘girls’ lab». 

Connie St Louis quoted extracts of Sir Tim Hunt’s speech augmented by her own comments which condemned what she perceived as a sexist comment\(^4\). Her tweet was retweeted more than 600 times, triggering a violent social media campaign which was amplified by the mainstream media and resulted in the destruction of Sir Tim Hunt’s lifelong career in less than a week.

**2. Study design**

In order to understand the dynamics of this controversy and how it led to the outcome we all know, we used both qualitative and quantitative methods. We first studied the activity generated by this controversy on Twitter by extracting several attributes from tweets such as the author, the text, the date, the language, the prominent hashtags and the mentions. We then plotted the distribution of tweets on this controversy over time in order to observe the rise and decline of the controversy (§ 3.1). We next sought to determine the sentiments expressed in the tweets by calculating the polarity of their constituent words (§ 3.2).

---

Thirdly, we tried to identify who the main protagonists were by generating the citation network of tweets (§ 3.3). Finally, in § 3.4, we looked at the digital footprints of the main protagonists on the web in order to understand why they were so successful in swaying public opinion on social media, thus resulting in the downfall of the “accused” and his removal from all his honorary positions (University College London and the Royal Society of England).

3. Analysis of traces of the controversy on Twitter

Because it is practically impossible to obtain all tweets on a topic a posteriori from Twitter, we searched a sample of 11 million tweets which represented 1% of all tweets produced in June 2015. Considering the large number of tweets searched, our observations should remain valid with respect to the tweet distribution. From these 11 million, we extracted those tweets containing hashtags and mentions related to the controversy. By decreasing order of frequency, they were “#distractinglysexy (1705 tweets); #feminism (952); #timhunt (795); @royal_society (152); #sexism (138), @connie_stlouis” (77), although the hashtag #sexism showed an even distribution over the period studied (June 2015) and is therefore not only related to this specific controversy.

3.1. Distribution of tweets related to the controversy

The majority of the tweets were in English: 84% of tweets for #distractinglysexy and 87% of tweets for #timhunt, respectively. This is not surprising given that the controversy concerned a British scientist and took place in an international setting. However, echoes were found in other languages although at a significantly lower percentage: 7% of tweets with the hashtag #distractinglysexy were in Spanish and 1.4% in Korean. Only 4.3% of tweets with #timhunt were in Spanish. This percentage was significantly lower in Korean with only 0.3%.

Figure 1 hereafter plotted the distribution of tweets with the three major keywords “#timhunt”, “@connie_stlouis” and “#distractinglysexy”. The latter was a hashtag coined by the feminist VagendaMagazine to launch memes that appeared two days after the start of the controversy. “#distractinglysexy was used to deride Tim Hunt’s portrayal of women as being emotional in science labs. The frequency of this hashtag rose from 28 on its first day to 449 on its second day and then to 700 occurrences on its third day. In less than a week (between 8th-16th June 2015), most of the keywords linked to this controversy witnessed a sharp drop in frequency on Twitter. This sharp drop corresponded to the resolution of the controversy after Sir Tim Hunt was publicly disgraced and forced to resign from all his prestigious positions in the real world.
Figure 1: The distribution of the three major keywords related to the controversy between the 9th of June and the 1st of July 2015

Figure 2 hereafter shows the word clouds generated from the tweets with the hashtags “#distractinglysexy” and “#timhunt”, respectively. The size of a word is proportional to its frequency. Unsurprisingly, in tweets with #timhunt (Figure 2a), the hashtag “#distractinglysexy” stood out as the most prominent since it contained the memes launched by the feminist “VagendaMagazine” deriding the scientist’s perceived sexist views. This hashtag also had the highest frequency in the period of the controversy (see Figure 1 above). Other frequent words found in the #timhunt cloud were “http” and “rt”. While these are low semantic bearing words, they are actually quite informative in the context of social media and 144-character universe of Twitter. Http indicates links to online content on the subject of the tweet, thus showing that the topic is receiving coverage in the mainstream press or in other online communities. “rt” denotes retweets of a tweet, thus is a sign of its propagation on social media. Other terms that were prominent in this word cloud were “scientists, womeninscience, female, crying, womeninstem” which are all expected given the topic of the controversy.
However, we notice the hashtag “reinstatetimhunt” in this cloud, referring to a counter campaign to reinstate the fallen Nobel Prize winner to his previous positions. The words in the “#distractinglysexy” cloud mostly relate to the memes mocking Sir Tim Hunt’s outdated view of women in science labs.

Figure 2: Tagcloud for tweets with hashtags #timhunt (left) and for tweets with #distractinglysexy (right)

3.2. Estimating the sentiment polarity of tweets

We used a supervised model of a sentiment polarity analysis tool trained on tweets from Sentiment140 called “echo” (Hamdam et al. 2015). Three positions are considered: negative, neutral and positive. The tool classified the majority of tweets with “#timhunt” as neutral (45%), while only 30% were classified as positive. These tweets were mostly announcing facts given the high frequency of the word “http” in them which is why they were classified as neutral. The majority of the tweets with “#distractinglysexy” were also classified as neutral (47%), while 34% were classified as positive. A closer look at the tweets showed, however, that they were largely ironic or sarcastic. Irony and sarcasm are notoriously difficult to detect automatically using classification algorithms because positive sounding words are employed in a negative sense. This aspect calls for further investigation.

3.3. The main propagators of the controversy and their network of influence

Next, we generated an author citation network of retweets and mentions at the peak period of the controversy (Figure 3). The size of a node in this graph indicates the

5 http://help.sentiment140.com/for-students.
number of retweets and mentions received by the person or institution represented by the node. It is therefore a good indicator of the person’s influence in the interaction network which led to the resolution of the controversy.

The central role played by “@connie_stlouis” in triggering the controversy is apparent in this network as well as that of the “@royalsociety” which was at the receiving end of the social media campaign to remove Tim Hunt from its membership. The role played by four other protagonists, namely @Carlzimmer, @2casey451, @Deborahblum and @Docfreeride was also revealed by this graph. In fact, the controversy began as a conversation between @connie_stlouis and the four protagonists before it reached other users.

Figure 3. The citation network of authors of tweets between 9-11th June 2015

3.4. Digital footprints and social status of the main propagators

To understand the role these prominent propagators played in constructing the narrative about the object of the controversy (Sir Tim Hunt), we searched the Internet for their digital footprints in order to determine their social capital as defined by Pierre
We focused on the largest nodes in the citation network (minus @Royalsociety which is an institution). We found that the majority of the main propagators (9 out of 12) were people with a high intellectual and social capital in the real world. Most were women scientists, journalists and writers who are actively engaged in fighting gender stereotypes. These influential Twitter account holders are well versed in the use of social media to advance their causes. All had a significant number of “followers”.

However, none of the main propagators had reached the level of social capital and notoriety of the accused (Sir Tim Hunt) who is a Nobel prize winner and a world-leading scientist in his field. Of further importance, other leading scientists including eight other Nobel prize winners came forward in the wake of the public disgrace of Sir Tim Hunt and signed petitions asking for him to be reinstated in his official roles based on his more than 50 years of real-life unblemished scientific record rather than on his unfortunate and careless utterances at the end of a wine-laden lunch for which he has profusely apologised. Furthermore, it has since emerged that his principal accuser, the adjunct lecturer and journalist Connie St Louis had exaggerated her qualifications and work experience in her CV\(^7\), thus raising questions about the credibility of her account of Sir Tim Hunt’s utterances which some have found one-sided. All the latter facts that emerged to give a more balanced and truthful account of this controversy have been ignored and Sir Tim Hunt remains disgraced socially and professionally. Therefore, it did not seem to matter that the main propagators were not the most knowledgeable or “famous” people with a big social capital, it sufficed for them to have a forum on the web (bloggers or web sites, social media accounts), to hold extreme opinions (feminist, anti-racist, extreme right wing or left wing, etc.) and for them to have sufficient “followers” to propagate their opinions, leading to an echo chamber in the mainstream media and subsequently to public institutions making decisions under the pressure of social medial campaigns.

4. Future work

Ours is a preliminary study on how social media have become the “public arena” on which important issues are debated and resolved, leading to real-life consequences that challenge how knowledge and truth are elaborated in the 21st century. It would seem that how knowledge and truth are arrived at may well depend not so much on who you are in real life and your ability to put forward verifiable facts but on your social media

\(^6\) «Social capital is the sum of the resources, actual or virtual, that accrue to an individual or a group by virtue of possessing a durable network of more or less institutionalized relationships of mutual acquaintance and recognition. (Bourdieu, in Bourdieu and Wacquant 1992: 119)» cited in David Gauntlett, Accessible at http://www.makingisconnecting.org/gauntlett2011-extract-sc.pdf. Accessed on 27/01/2018.

\(^7\) For a full account of the post-controversy analysis bringing to light many troubling facts about Connie St Louis can be found here: http://www.dailymail.co.uk/news/article-3141158/A-flawed-accuser-Investigation-academichounded-Nobel-Prize-winning-scientist-job-reveals-troubling-questions-testimony.html#.unRK0d7kk.
network and activism in the virtual world. Further study is ongoing on other controversies also propagated on social media in order to determine if they share the same attributes. Some of the questions we seek to find answers to are: do most social media controversies have a very short life span (a matter of days) before resolution, as opposed to controversies in the real world which can last several weeks, months or even years? What are the profiles of the main protagonists and propagators? What impact do these controversies born on social media have on real life organisations and people?

Ibekwe-SanJuan and Bowker (2017) observed that “ongoing transformations in knowledge production processes entailed by Big Data and web 2.0 put pressure on the KO community to rethink the standpoint from which Knowledge Organization Systems (KOSs) are designed. (...) Theoretically, this entails a shift from purely universalist and normative top-down approaches to more descriptive bottom-up approaches that can be inclusive of diverse viewpoints”. How can this dilemma of enabling both top-down and bottom-up modalities of knowledge production be resolved in the era where “post-truth” and “alternative facts” are being propagated without jeopardising the integrity of KOs? This will call for new philosophical and methodological approaches to KOs that are adaptable and flexible, involving humans and machines working in real time to resolve contradictions.

References
Copyright infringement: between ethical use and legal use of information

Abstract
The purpose of this communication is to highlight the singularities that permeate the ethical and legal dimensions of the use of information, revealing some disagreements between both in order to contribute to a critical view of the use of information. The study reveals how ethical and legal dimensions are often treated synonymously, taking as an example some dilemmas concerning the differences between property rights and moral rights to demonstrate how an ethical issue cannot be reduced to a legal issue. The conclusion of the research highlights the importance of the intervention of librarians to train users in the ethical use of information.

Introduction
Alexandra Elbakyan, a Russian neuroscientist from Kazakhstan, was only 20 years old in 2009, when she was writing her thesis on biometric scanning for consumer electronics. Like any other Master’s or PHD student, she needed access to other people’s research to finish her own, but the problem is that most of the world’s scientific knowledge is owned by online publishers such as Elsevier, Springer, Wiley-Blackwell, Sage and Taylor & Francis. Together, they publish around half of all scientific papers produced in universities around the world, and they do so by charging authors to publish their work and readers to read it.

When Elbakyan realized she had to spend around 30 dollars to have access to each paper – and she needed access to dozens of them – her decision was to go to the internet after pirated copies of those publications. Two years later, she used her computer skills to create Sci-Hub, a website that bypasses publisher paywalls by allowing access to scientific papers through educational institution proxies. The papers downloaded by Sci-Hub are stored both in its own repository and in Library Genesis (LibGen), making over 64 million academic papers and articles available for direct download without the need to pay for them.

The question we ask here is: can something illegal be considered ethically justified? On the other hand, is it possible that something that is acceptable according to the law is open to debate when it comes to its ethical validity? Thinking about these questions might lead us to the understanding that the ethical and legal uses of information do not always accord with each other.

In this paper, we intend to highlight the singularities that permeate the ethical and legal dimensions of the use of information, revealing some disagreements between both in order to contribute to a critical view of the use of information. The study reveals how
ethical and legal dimensions are often treated synonymously, as we see in the documents concerning the definitions of information literacy published by The Association of College and Research Libraries (ACRL) and used as reference by many information researchers around the world.

Following that discussion, we take as an example the dilemmas concerning copyright laws in digital networks to demonstrate how an ethical issue cannot be reduced to a legal issue. Finally, we focus on the call for the intervention of librarians to train users in the ethical use of information.

Ethics and the law: are they the same thing?

The concern that motivated this research, as mentioned above, came from a search for the use of the "ethical" and "legal" predicates in the two main ACRL documents on information literacy: Information Literacy Competency Standards for Higher Education, published in 2000, and Framework for Information Literacy for Higher Education, published in 2015. In both Standards and Framework, the term “ethics” is used in only three passages (in the Framework there are four, but one passage is repeated). In the first document, one reads that, among other attributes, an information-literate individual is able to “understand the economic, legal, and social issues surrounding the use of information, and access and use information ethically and legally” (ACRL, 2000, online); in the second passage, the document repeats that the information-literate student “accesses and uses information ethically and legally”, attesting that this student “understands many of the ethical, legal and socio-economic issues surrounding information and information technology” (ACRL, 2000, online).

We see that the legal and ethical dimensions always appear next to each other, practically synonymously, without much of a conceptual distinction between them – or even an explanation as to why they seem to be, basically, the same thing.

The Framework was published fifteen years later as a way of updating the old one, in an attempt to incorporate the criticism held against the technicality of the previous standards – now seen as a framework. In this new document, the old definition of information literacy as a set of abilities requiring individuals to “recognize when information is needed and have the ability to locate, evaluate, and use effectively the needed information” (ACRL 2000, online) was updated to become “the set of integrated abilities encompassing the reflective discovery of information, the understanding of how information is produced and valued, and the use of information in creating new knowledge and participating ethically in communities of learning” (ACRL 2015, p. 3). In the Framework, the term ethical emerges as an adverb regarding the way people participate in their “communities of learning”. But we still cannot grasp what kind of participation would be considered “ethical”.

Throughout the document the term appears two more times: in the introduction, we
read that “students have a greater role and responsibility in creating new knowledge, in understanding the contours and the changing dynamics of the world of information, and in using information, data, and scholarship ethically”; finally, when the dispositions for “learners who are developing their information literate” are listed, it is written: “follow ethical and legal guidelines in gathering and using information” (ACRL 2015, p. 2). Thus, even though the document brings new and wider views about information literacy, it still attaches the idea of an ethical use of information to a legal use of it.

Copyright and piracy

With more than 400,000 articles published every year, Elsevier has a high profit margin in its business approaching nearly 40% - which amounted to more than US$2 billion in 2016. The company obviously got upset with Elbakyan’s initiative of giving free access to scientific papers. As a result of a lawsuit brought by the company, the New York district court awarded Elsevier US$15 million in damages for copyright infringement by Sci-Hub, the Library of Genesis (LibGen) project and related sites. Elbakyan defends herself by making reference to Article 27 of the United Nations Declaration of Human Rights: “everyone has the right to freely participate in the cultural life of the community, to enjoy the arts and to share in scientific advancement and its benefits”. According to structural biologist Stephen Curry at Imperial College London in the United Kingdom, “Sci-Hub is obviously illegal (…). But the fact that it is so immensely popular, inside and outside academia, is a symptom of many people’s frustration with the status quo in academic publishing”1.

Sci-hub is described in its virtual page as “the first pirate website in the world to provide mass and public access to tens of millions of research papers”. The word “pirate” comes from the fact that the website violates copyright laws by providing free access to content that is typically charged by publishers. Copyright infringement, in this case, relates to property rights and not to moral rights: the latter are related to authorship, and differ from those for being "inalienable, which implies their characterization as personality rights, because they are inherent to the subject of law and immune to any transaction or subtraction" (Bezerra 2014, p. 102, our translation). Therefore, they do not involve the economic relationship established between the creator of the work and the company (in this case, the publisher) responsible for its commercial exploitation.

Although the violation of both property rights and moral rights ultimately characterizes a violation of copyright laws, it is possible to realize that the ethical assessment of both is not the same, and that the moral judgment of those who "pirate" a scientific article to expand access to the knowledge contained therein is distinct from

the judgment of someone who incurs plagiarism or false attribution of authorship to garner some kind of benefit.

**The role of librarians**

Since the turn of the century, problems with truth and fallacy on the internet have become a subject of debate. Forbes and Mintz (2002) state that the internet has become a dangerous sea to navigate because of the huge amount of false truths, misinformation, misleading or erroneous information, under a myriad of shapes and guises. Given internet users’ neglect, innocence or lack of preparation, the authors warn about the dangers that imprecise data may trigger, namely regarding their health, privacy, investments, business decisions, online shopping and legal matters. Mintz (2012) summed up the subject, stating that new lies and camouflages are being used to scam unsuspecting users, namely digital identity theft, appeals to hate and religious and racial intolerance, ecommerce frauds, information warfare and cybersecurity, political shenanigans, charity scams, and so on, referring that it is more and more important to adopt preventive behaviour in response to such risks.

In the academic environment, the search for information on the internet is an inescapable need, so a risky digital context is a particularly sensitive issue (Jasanoff 2016). At the same time, we are living in an era where any internet user can also produce contents and disseminate information. So the issues raised in educational institutions are very wide-ranging: people must be empowered to surf the internet safely, but it is also essential that they be taught to filter the contents that wish to use or post and make available to others. Digital autonomy and freedom on an individual level imply greater civic and social responsibility, elevating the importance of ethical and legal matters associated with searching, using and releasing information.

In an original study, Saraiva (2016) gives a detailed explanation of fraud and deceit in the academic context, suggesting a behavioural analysis together with a legal and economic perspective of the situation. Giving examples from the literary universe of Harry Potter, she shows how in several situations the academic context is a sanctuary for improper, immoral or illegal conduct. Based on an in-depth analysis, the author mentions that correcting this kind of behaviour (which crosscuts teachers, researchers and students) implies knowing what is behind it, and this involves identifying the costs and benefits underlying the deception and the risk taken, from an individual and collective perspective. Finally, as a solution she points to involving teaching institutions in the prevention and correction of such behaviour. For, as the author states, “as institutions serve as behavioural anchors, the readjustment of a new choice architecture implies, surely, its re-appreciation (at least in relative terms) and a re-equation of values, mainly academic integrity” (Saraiva 2016, p. 343). The main answer lies in pedagogical training for academic integrity. However, this should take into account that ethics is flexible and dynamic or, in other words, situational, leading to behaviours that
adapt, leading to different moral judgments in the face of similar dishonest behaviour. Hence the importance of the legal, fixed aspects becoming a deterrent, though they cannot be the only tool. Students’ proximity, trust, training in the right skills, a peaceful environment that supplies food for thought, all discourage cheating, because they help internalise academic rules, gain moral self-perception and promote a feeling of belonging. All these factors contribute to a framework that implicitly disciplines behaviour.

Regarding students’ behaviour towards information, Harris (2011) defends the importance of damage control with respect to the above mentioned risks. He suggests that in order to deal with discipline, a compromise of solutions from three approaches must be found: regulatory, technological and pedagogical. The first approach includes legal and regulatory issues. The second deals with computer security matters such as anti-virus, firewalls and other cyber-security tools. The pedagogical approach includes the proper use of information found on the internet, namely in content assessment with criteria such as purpose, authority, truth deconstruction techniques (fact verification) and education for the use of different analysis and evaluation tools: “The goal is to make thinking about evaluation automatic, a habit of mind that is deeply ingrained into the process of Information use in any context” (Harris 2011, p. 176).

Several authors consulted agree on one point: in a context where social and academic pressures on students call for better results and, at the same time, competitiveness, the need for lifelong learning and the requirements of multiple skills are evaluated, the support and guidance provided to students, specially concerning work in the library, must be a priority (Silva, Jambeiro, Lima and Brandão 2005; Harris 2005; McAdoo 2012; Heine and O’Connor 2013; Lopes, Sanches, Andrade, Antunes and Arévalo 2016; Ramírez Leyva 2017). Therefore, training in information literacy is particularly useful. As a direct consequence of the development of skills in searching, locating, selecting and using information, there are tangible benefits for students. Information literacy skills are an indispensable investment towards a future of participative citizenship.

The imperative to train students in these skills is illustrated in the Framework for Information Literacy for Higher Education (ACRL 2015). It stresses dynamism, flexibility, individual growth and community learning, organized around six key concepts, directed towards the understanding and application of information literacy: Authority Is Constructed and Contextual; Information Creation as a Process; Information Has Value; Research as Inquiry; Scholarship as Conversation; Searching as Strategic Exploration.

In particular, when information is referred to as a value, many of the ideas at stake regarding this topic emerge, in the expectation that information-literate students know how to (ACRL 2015, p.6) give credit to the original ideas of others through proper attribution and citation; understand issues like intellectual property, copyright, fair use,
open access, and public domain. The same document underlines that students who are developing information skills should (ACRL 2015, p.6): respect the original ideas of others; value the skills, time, and effort needed to produce knowledge; see themselves as contributors to the information marketplace rather than only consumers of it; are inclined to examine their own information privilege.

Allan (2010) stressed the importance of developing skills in information literacy in second and third cycle students of higher education, when the need to handle information is accentuated. At the beginning, a very basic range of sources of information and a simple internet search are used. However, in order to succeed, students need to develop advanced skills in searching and managing information. This is the phase when librarians should take action (Allan 2010, p. 97). Hence the importance of educational provision in matters regarding citations and references, taught by librarians. Lopes (2013, p. 23-24) underlines the importance of citations for the reader to confirm several situations: that the text was read, that the information obtained therein was incorporated in the construction of our knowledge and that there is an overall commitment to the contents and authors cited. Similarly, other authors (Burbules 2015; Alonso-Arévalo 2017) state that citations are not only a manner of sustaining the author’s opinion, but also a demonstration of the networks created among researchers, based on recognition and respect, and an important way of evaluating scientific production within a complex system that manages bibliometrics and impact factors. Librarians are parties involved in the ethical use of information. The fight against plagiarism and fraud should stem from an ethical culture, where respect, the establishment and transmission of principles regarding authorship, right to privacy, and social responsibility, among others, may be a meaningful way of changing students’ behaviour in this regard (Bodi 1998; Puga 2014).

Conclusion

We understand that ethical and legal dimensions are not the same thing; moreover, we see that much of the legal field is built around values and practices that a society develops within its culture (given all the power struggles that accompany this process). The conclusion of the research is that the moral judgment that permeates new cultural practices tends to be consolidated before being contemplated by the legal system, and that, therefore, the ethical use of information will not always be in agreement with its legal use.

In the digital environment, where many new practices and possibilities emerge very rapidly to be legally examined and contemplated, it is common for certain acts to be considered ethically positive (although some of them may be legally prohibited) as well as the opposite. For better or for worse, the legal system tends to act after certain social practices spread, so it is very important to act in the context in which these practices occur, namely in an academic environment, and discuss these practices in a way so as
to make the population familiar with the new regulations which, in the end, will shape their life practices.

In conclusion, we feel it is possible to infuse an ethical component into the teaching of information skills. The implementation of information literacy, considering the results presented, urges librarians to perform a more comprehensive task, raising students’ awareness about authorship possibilities without interfering in ethical matters. In this sense, our investigation has sought to contribute to the understanding of information literacy as a conceptual framework for shaping the experience of school work in academic libraries with the essential ethical component that students can develop in this context.

References


Dealing with the paradoxes of customer opinion for effective decision support in churn management

Abstract
The process of analysing opinion expressed about a brand, organization or product in churn management has been taking on a new dimension in recent times. A subjective opinion, which hitherto could be rendered as positive or negative, can now be positive and negative. Accordingly, the act of channelling an effective acquisition or retention strategy for customers becomes more complex with increasing overhead costs. Moreover, the increase in false positive and negative churn classification via sentiment analysis affects organizational knowledge, which oftentimes causes other societal challenges. As churn prediction analysis transits dramatically from local transactional data analysis, to social network content analysis for real time churn decision support, the paradoxical nature of opinion may render sentiment analysis as a tool for opinion mining less effective for decision support. This is because the sentiment analysis approach handles customers’ opinion as an independent entity; while oftentimes, user opinions are used only once for defining the behavioural class of a customer. These processes neglect the relative exclusive association, and the influence of community members over one another in churn management. This paper presents a context-based clustered conversation model for churn complexity management via index-driven opinion within a community of social network users. The essential aim is to cluster opinion about knowledge organization that is scalable, incremental and contextual for an appropriate churn management decision support tool.

1. Introduction
In today’s competitive market, knowledge and the capacity to create and utilize knowledge are considered the most important source of a firm’s competitive advantage (Hosseini et al. 2012). However, many companies are unable to function as knowledge-based organizations because they suffer from learning disabilities, and cannot create, identify, and value their knowledge assets (Lai 2007). In a world of dynamic change, with the influence of the Internet on everything from marketing and customers to knowledge discovery, the pace at which organizations are constantly seeking ways to adapt themselves to new conditions in order to survive and flourish in a competitive marketplace cannot be underestimated (Chi 2011). Through several micro blogging platforms like Facebook, Twitter (Gruzd et al. 2011), the volume of unstructured data keeps increasing everyday with users expressing their opinion on different issues at different times and for different purposes (Gupta 2014). Twitter, with its vast number of tweets, and millions being added every day, serves as a worthy platform for sentiment analysis due to its large user base drawn from a variety of social and cultural regions worldwide, but aside from this fact, it can also be used for prediction (Liu et al. 2011). Such clusters of unstructured data consist of a gathering of qualities and variables related in a particular sense and varying in another. Most organizations typically use this data once and then lose it, despite the savings they could realize by
reusing it. With this in mind, some have concluded that it is necessary to view knowledge organization through the lens of big data attributes as well as through the functional goals of the user/customer (Michael and Carolyn 2013). In addition, a key to maximizing the values of decision support in churn management through knowledge organization is to be able to detect potential churners before they leave the service and target them exclusively for such campaigns. No doubt, churn prediction analysis (Dasgupta et al. 2008) towards effective customer acquisition or retention strategy has become an integral part of business strategy. In customer relationship management, a current business approach, which used to be managed intrinsically (Yan et al. 2004) through customer profiles and/or inherent features of the service, is also now being extrinsically managed to portray the service in terms of the value it accrues through its social roles (Karnstedt et al. 2011). This is necessary because humans are influenced by the actions of others in the society. Thence, the idea of social network analysis (SNA) is to pay close attention to the actions of an immediate social circle through behavioural analysis (Richter et al. 2010). Consequently, one of the biggest challenge will be a matter of turning big data into knowledge discovery, an insight that will enable organizations to make cogent decisions. So far, the methods employed via sentiment analysis have been less effective due to the independent analysis of each opinion without a communal influence. While conflicting, noisy, and missing data (Holzinger et al. 2013), as found in the increasing volume of unstructured data, is also a factor, the paradoxical nature of opinions are mostly not considered by the existing sentiment analysis approach. The paradoxical attributes of an opinion show that feelings about a brand sometimes seems self-contradictory or absurd but in reality may also express a possible truth, which can be complex. Thus, mere classification of sentiments inherent in this opinion can be vague, such that optimal decisions derived from respective churn class may not be accurate because the complexities in processing customers’ available information can also produce more complexities in decision support. This, process leads to an increase in false positive (FP) and false negative (FN) for various customer churn classification analyses conducted via social media other than feature-based approaches. Solving this problem requires an approach to organizing and indexing contents (opinions) for a sustainable period, to handle the complexities in decision making for respective churn classes. In addition, for the complexities to be adequately managed, effective decision support paradigms for potential churners must be put in place. Here, a scalable, incremental, contextual clustered conversation (SICC) algorithm is proposed to cluster organized opinions about a product, or brand of a company, over a period. Accordingly, the organised indexed tweet is clustered based on context, response(s) and influence on the members of the community. The essence of the context-based conversation cluster of opinions is to enable the almost real-time determination and classification of customers’ churn class via the customers’ affinity network analysis and expressed opinion. The churn class of a customer informs the
decision to be taken by the organization at a given time. Therefore, in section 2, a literature review on related works is presented, while in section 3, the research motivation is discussed and the proposed context-based conversation cluster is presented in section 4. In section 5, sample experiment and evaluation on users’ experience and response expressing opinions about product or brand of an organization is presented using a sample of 1,350 active tweeter account handlers. Finally, the research is concluded in section 6 with a brief overview of future research intentions.

2. Literature review

Since the emergence of social media on the Internet as an important evolving knowledge asset (Apoorv et al. 2011), the field of customer relationship management (CRM) and marketing through these platforms has created many opportunities by attracting customers with engaging content and creating communities to leverage their products. This rapid development of digital systems and associated information technologies has provided a reliable digital CRM system, which has begun to enhance opportunities to understand customers (Xu 2011) through customer churn management. Historically, CRM and knowledge management initiatives have been directed towards the delivery of continuous improvements for customers (Gebert et al. 2003) to build profitable, long-term relationships. However, to achieve the knowledge-enabled customer relationship management goal, an effective churn prediction must be in place.

Customer churn, which is often referred to as customer attrition or customer turnover, is the loss of existing customers to another company or service provider (Kerdprasop et al. 2013). This process has been extensively studied in various domains such as banking, employee management (Saradhi and Palshikar 2011), online gaming, airlines, and telecommunication services (Hossain and Suchy 2013) among others. Based on this, various data mining technologies, such as Clustering (Popović and Bašić 2009), Decision Tree (Rahul et al. 2011), Neural Networks (Kamalraj and Malathi 2013), Regression (Amin et al. 2014), Support Vector Machine (Kim et al. 2012) and an ensemble of hybrid methods (Kumar and Ravi 2008) have been used for churn predictions using intrinsic customer features such as demography, age, pricing, service failure rate, etc. For a successful data mining process to take place, the inherent noise in the dataset must be well taken care of. Feature selection (Verónica et al. 2011) is a process of removing the irrelevant and redundant features from a dataset in order to improve the performance of its learning algorithm. The approach (Ibitoye and Onifade 2016) is required to filter and select the most optimal data subset that is required for churn prediction management. No doubt, the result of a churn prediction model is largely dependent on the success of the future selection model. While the intrinsic factors used in churn prediction do not account for the role of social ties between individuals in affecting the propensity to churn, the extrinsic factors (Verbeke et al. 2011) portray churn prediction service in terms of the value it accrues through its social
roles e.g. community opinion, effect of word of mouth. The idea that similar people tend to interact more frequently, otherwise known as homophily, offers the possibility to predict a person’s behaviour based on the observed behaviour around them (Zhang et al. 2012). One great platform where this may be achievable is through social media via social network analysis (Shuang-Hong et al. 2014). This is because social media websites provide a public forum that gives individual consumers their own voice, as well as access to product information that facilitates their purchase decisions (Kozinets et al. 2010). Examples of social media that are popular among all level of consumers include blogs, YouTube, MySpace, Facebook (Sin et al. 2012). However, based on its scalability, and accessibility to all, Twitter, a fast growing online social network, allows users to upload short text messages, also known as tweets, with up to 140 characters. It is often used for customer behavioural analysis (Liu et al. 2011) conducted for effective customer relationship management. Consumer behaviour is the study of individuals, groups or organizations about their process of selecting, securing, using and disposing products, services, experiences or ideas, to satisfy needs and the impact of these processes on the consumer and society (Nan-chan 2009). Since customer behaviour modelling has gained increasing attention in the operation management community (Shen 2007), especially via extrinsic factors, a significant field of study that analyses people’s attitudes towards entities, individuals, organizations, products, services, events, and topics, and their attributes (Liu 2011), is referred to as opinion mining or sentiment analysis. It is a well-suited process for various types of business intelligence (Shahheidari et al. 2013) while its timely manner, at a low cost, do have a great influence on customer perceptions and behaviours. The process of building a system to collect and examine opinions can be applied at four different levels; (Tan et al. 2011) the document level, sentence level, aspect level and the level of user or social relationships between different users. These levels can be adequately processed through natural-language processing, (NLP) lexicon or machine learning solutions, by labelling text manually or through a hybrid model, which requires the effort of human experts in labelling and training the data sets. While the machine learning-based approaches for opinion mining are supervised learning tasks, which utilise textual feature representation coupled with classification algorithms to infer the opinions expressed in the text (Adam et al. 2015), the lexicon-based technique is an unsupervised approach, which relies on the assumption that the collective polarity of a sentence is the sum of the polarities of the individual words or phrases that constitute it. (Chetan 2014) The objectives of these methods are to define positive, negative, or neutral feeling through text (Medhat et al. 2014) as a tool for customer retention or acquisition incentives. However, when such personalized marketing is implemented in CRM strategies as a sole entity without the influence of the community, the value of the knowledge of the social media structure cannot be fully realized. Thus, majority of incentives have a huge potential to be wrongly targeted at customers. Therefore, the need for an organized
knowledge flow from the company to the customer is a prerequisite. No doubt, when knowledge management is being upheld by information technology, organizational performance through enhanced efficiency will greatly improve. However, for an organization to recognize the requirements and expectations of customers through customer knowledge management, organizations must cluster carefully and process the customers’ implicit knowledge of their brand or product as represented through expressed opinion on social media, to find a better understanding about customer needs and expectations. This organized knowledge for effective decision support in churn prediction is presented in section 4 as a contextual clustered conversation model.

3. Research motivation

As events unfold daily on microblogs, they display real-world occurrences over space and time. Although exchanging information is not a new concept on the Internet, much of the information exchanged on social networking sites has always been conversational in nature. However, its particular importance to churn management has not been adequately explored towards effective knowledge organization. Here, we define conversation retrieval as a subset of event detection paradigms that intercept information retrieval and social network analysis. Over the years, several existing conversation-clustering algorithms have focused on the mere use of packages of words through a single pass incremental model that does not cater for evolving tweets with respect to initial conversations. Moreover, the methods are not scalable enough to identify multiple contexts through multiple conversation clustering detection analysis. Thus, extracting meaningful information is a major challenge of tweet mining. At the same time, attempts by several sentiment analysis approaches have been more complex and less effective due to the paradoxical nature of opinion. In solving these problems, a context-driven conversation clustered around indexed pre-processed tweets is presented in this research. The process of achieving this goal is further illustrated in section 4 below.

4. Clustering Tweet conversations based on social context

In order to engender organised knowledge; using unstructured user opinions, which are generated in real time via social network analysis, in this research, a semi-supervised algorithm for event monitoring is used to build a structured cluster of 1 to n conversations. The proposed organized knowledge for opinions as illustrated in figure 1 below is contextual (based on type), incremental (no time limit) and scalable (functions better when volume increases) with and without any prior knowledge.
Figure 1: Organized Knowledge through clustered conversation

From figure 1, $t_{si}$ is the tweet streaming at time $t_i$, while $TC_{ci}$ is the tweet conversation class identifier. At first, tweets are pre-processed in order to clean tweets from noisy tokens and characters, normalize the language used, and generalize the vocabulary used for the expression. Thereafter, the tweet words are indexed to keep track of the tweet for onward processing. Since tweets are represented as word vector spaces, during the process of clustering conversation, a word can belong to one or more conversation/document vector spaces, the goal of indexing is to enhance the contextual semantics of tweets through direct ordering analysis. Then, a scalable, increment and contextual conversation (SICC) algorithm begins by finding the root of a conversation for the text clustering processes. Usually, the first tweet on a subject matter defines the root of a conversation among a given set of social network users. Here, the inherent special features, like the hashtag (#), @reply, mention (@username), are used to determine the context of tweets on arrival while other text through a word vector space decomposition is used to enhance the contextual clusters. Therefore, we define a set of context labels based around company name, company products, product prices, product brand, and product qualities, among others, in order to generate multiple contexts of conversation. The algorithm listing to achieving this goal is explained briefly in table 1 below:
Algorithm Listing 1: Scalable and Incremental Conversation Clustering algorithm for Tweets

Input: Let $prt_i$ be the indexed Pre-processed Tweet
Output: Context Based Conversation Clusters.

1. Let $C_i$ (where $C_i$ is from 1 to n) be a set of clustered conversation
2. Let $prt_0$ be the first tweet (root) published in the clustered conversation $C_i$
3. while (type ($prt_i$) != root) do
4.   For every $prt_i$:
5.      Extract # from $prt_i$ by matching its field in relative to $C_i$ status. (if any)
6.      Extract @reply from $prt_i$ by matching its field in relative to $C_i$ status. (if any)
7.      Extract mention from $prt_i$ by matching its field in relative to $C_i$ status. (if any)
8.      Extract other features in $prt_i$ then
9.      Find the conservation sequence of $prt_i$ using Hidden Markov Model
10.     Add $prt_i$ text to cluster $C_i$ while $ST_i$ is the set of all words in $C_i$
11.    End for
12.  End while

The above algorithm in listing 1 goes beyond contextual classification of tweet with hashtag to defining the context of the clustered conversation by finding the social relationship between users through the @mention and @reply (RT) functions. Since long conversations are scarce on Twitter, the algorithm process also monitors the sequence of the dialogue by using the Hidden Markov Model (HMM). This is necessary because every user on a social network like Twitter is entitled to their opinion. Thereby, making information drift from ongoing conversation to starting a new discussion is the norm on the platform. Hence, to have a robust set of contextual conversation cluster, for every incoming tweet, the system checks and extracts it hashtag (if any), @reply (if any) and a mention (if any) to determine if a relevant conversation cluster(s) to the tweet is in existence. If a clustered context is already defined for the parameters obtained from the incoming tweet, the HMM function is used to determine the tweet sequence level by measuring its structural dependencies between existing utterances in various clusters. Thereafter, the tweet is added to the most appropriate conversation cluster using the probability function in equation (1) below.

$$p(C_i|T_i) \propto p(C_i|cT_{i-1}) \cdot p(C_i|T_{i+1}) \cdot p(T_i|C_i)$$ (1)

While $C_i$ is the cluster of a tweet $T_i$, $T_{i-1}$ denotes the previous tweets while $T_{i+1}$ is the next tweet in a given cluster $C_i$. $\propto$ is the hyper-parameter that determines how likely a new cluster is created and $p(T_i|C_i)$ is the probability that $T_i$ is generated from $C_i$. Thus, from equation 1,

$$p(T_i|C_i) = \sum_{x \epsilon W(x|C_i)} count(t_i, x)$$ (2)

Where $x$ are the selected features used for context definition, $W$ is the entire word vector present in a tweet. $count(t_i, x)$ is a function that returns the number of occurrences of a feature $x$ for a tweet or a cluster. Then, from equation 2,
\[ p(x|C_i) = \frac{\text{count}(C_i, x) + \alpha}{\sum_{x \in W} \text{count}(C_i, x) + |W| \cdot \alpha} \]  

(3)

The essence of using this approach is to derive concise clusters of tweets that are contextually related and trackable as the volume of tweet increases over time. Oftentimes, a transition probability \( p(C_n|C_i) \) is required in engendering a new cluster. This is defined using the equation (4) below.

\[ p(C_n|C_i) = \frac{\text{transitions}(C_i, C_n) + \gamma}{\sum_{x=1}^{n} \text{transitions}(C_i, C_x) + n \cdot \gamma + \alpha} \]  

(4)

Where \( n \) is the number of occupied clusters, \( \gamma \) is the flooring value to avoid zero probability, the transition \( (C_i, C_n) \) returns the number of transitions from \( C_i \) to \( C_n \). Thus, a new conversation cluster \( (C_i = \text{new}) \) is generated if any of the extracted parameters do not match any of the available context conversation cluster using the probability function in equation 5 below.

\[ p(C_{\text{new}}|C_{T_{i-1}}) \cdot p(C_{T_{i+1}}|C_{\text{new}}) \cdot p(T_i|C_{\text{new}}) \]  

(5)

Where \( p(C_{\text{new}}|C_{T_{i-1}}) \) and \( p(C_{T_{i+1}}|C_{\text{new}}) \) are derived by using equation 6 below.

\[ p(C_{\text{new}}|C_{T_{i-1}}) = \frac{\alpha}{\sum_{x=1}^{n} \text{transition}(C_{T_{i-1}}, C_x) + \alpha} \]  

(6)

Based on these processes, an organized knowledge for contextual clusters of conversation is built from customers’ opinion on company brands and products. Based on this, appropriate decision support tools can be applied to target respective customers based on their churn class from expressed opinion mining. Therefore, to evaluate the need for the proposed organized opinion through social media (especially twitter) in churn prediction and decision support for customer relationship management, in section 5, an experiment on how well customers experiences in the use of social media have met their expectations is presented.

5. Experiments and evaluations

Based on a high percentage of active social media users, the following experiments were conducted with the goal of evaluating and establishing the need for organized knowledge in decision support for customer relationship management. Initially, we chose to find out much social media presence our respondents have. The results as presented in figure 2 reflect that over 40% have four social media accounts, 36.8% have just 3, 10.5% have 2, with 7.9% having just 1 account.
A further probe was to determine how often they use any of their social media accounts. A high percentage of our respondents use social media daily with few doing so only on a rare basis. Figure 3 further illustrates the values obtained.

Then, as part of a key factor to organizing opinions, we found out how frequently our respondents express their views about products or brands on social media. Quite a high percentage, as presented in figure 4 below, are passive on opinion expression on brands, i.e. they watch for a while before passing a comment.
Beyond mere expression of opinion, we also found out how often our respondents receive feedback from a company about their expressed opinions. This is important to determine the need for decision support in opinion clustering. The results are indicated in figure 5 below.

Figure 5: Feedback rating

Then, in figure 6, we chose to evaluate how our respondents feel if a company contacted them based on their expressed opinion. Then, in figure 7, we present their rate of satisfaction with the decision supplied by the company in response to their expressed opinion.
In figure 8, to establish the need of organized knowledge for effective decision support, we decided to discover how our respondents were influenced (if at all) by members of their social network in the choice of a brand or product. 46.8% of them are strongly influenced while the rest are presented below.
Through the developed organized contextual clustered conversations, in figure 9 we present the decision from the customers based on whether customers should decide the cost of a new data bundle before it is finally pronounced. 44.1% strongly agreed, 20% agreed, 3% are indifferent while 24.3% and 8.6% disagreed and strongly disagreed, respectively, as presented below.

Table 1 shows the features of each brand that the proposed algorithm was implemented on; in a bid to determine the overall decision rating of customers’ opinion on the product.
Table 1: Product features and description

<table>
<thead>
<tr>
<th>S/N</th>
<th>FEATURES</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bundle ID</td>
<td>The ID of the data bundle</td>
</tr>
<tr>
<td>2</td>
<td>Name</td>
<td>The name of the data bundle whether accepted or not</td>
</tr>
<tr>
<td>3</td>
<td>Size</td>
<td>The value of the data in megabyte if considered or not</td>
</tr>
<tr>
<td>4</td>
<td>Usage</td>
<td>Is it frequently used?</td>
</tr>
<tr>
<td>5</td>
<td>Price</td>
<td>Is the price moderate or not?</td>
</tr>
<tr>
<td>6</td>
<td>Preferred</td>
<td>Is it preferred?</td>
</tr>
<tr>
<td>7</td>
<td>Plan</td>
<td>Is it 24/7, weekend only, midnight or daytime plan?</td>
</tr>
<tr>
<td>8</td>
<td>Purchase</td>
<td>What is the frequency of purchased the data bundle</td>
</tr>
</tbody>
</table>

6. Conclusion
Since social network have massively enhanced globalization, communal beliefs, norms and values cannot be overlooked in customer relationship management. While customers are free to express their opinions about brands and products, companies are also investing hugely in how to make cost effective decisions. Several existing approaches to this process are not organized and do not consider community members’ opinions. However, they independently mine customer opinion to determine their sentiment class, which can be either positive or negative. In this research, an organised contextual clustered conversation of opinion is presented. The essential aim is to provide opinion clustering based on type, which can function better when volume increases and most importantly does not have a time limit) with and without any prior knowledge. The result of the experiments conducted further strengthens the need for organized knowledge to aid decision support in today’s customer relationship management.

References


Mariana Cantisani Padua, Maria José Vicentini Jorente, Natalia Nakano

Design Science as a methodology for the development of Knowledge Organization Systems in museological entities

Abstract
Information overload has brought marked changes in the way people search for information and, in turn, has reshaped society. Today, people want well-communicated, efficient information to help them accomplish tasks on the Internet, especially in digital information environments, such as digital libraries, digital repositories and digital museums. The aim of this article is to present a design method for creating a technological artifact, understood as a digital environment, hosting a digital museum that supports an optimized knowledge organization system based on human rather than subject-based aspects. It is an epistemological, bibliographic study, in which we suggest a theoretical reference based on Design Science Research for research and development of information systems with efficient knowledge organization in the digital environments of museological entities, and consequently efficient information retrieval. In general, the organizational structure of information appropriation in different contexts and media in digital environments of museological entities demands a re-contextualization of the museum object when it is transformed from the physical to the digital context. In this process, the interface is the locus where the interaction of the individual with the museum objects and the knowledge organization content of this digital environment take place. This interface should, therefore, be based on the principles of Design Science to provide quality solutions and innovations.

Introduction
Information overload has brought about marked changes in the way people search for information and, in turn, has reshaped society. Today, people want well-communicated, efficient information to help them accomplish tasks on the Internet, especially in digital information environments, such as digital libraries, digital repositories and digital museums. On the other hand, there is an exponential growth in the number of museum websites with diverse modes of knowledge and information organization based on traditional subject-based knowledge organization tools, which may be, at times, frustrating and fail to meet the demands of information society.

The aim of this article is to present a design method for creating a technological artifact, understood as a digital environment, hosting a digital museum that supports an optimized knowledge organization system based on human rather than subject-based aspects. Thus, the suggestion of this study is that digital museum environments be organized considering human aspects such as the reason for information search, cognitive functions and meaningful experiences, influence of social and technological aspects on knowledge growth, etc.

Method
This is an epistemological, bibliographic study, in which we suggest a theoretical reference based on Design Science Research for research and development of
information systems with efficient knowledge organization in the digital environments of museological entities, and consequently efficient information retrieval.

This study is justified since, according to the strategic map of museums, the issues related to preservation, diffusion and access, as well as social function, education and research comprise at least 50% of the museum's strategic actions (Ibram Management Report [IBRAM] 2016). Contemporary museology has faced challenges, leading to new study perspectives, without concern for specific themes and privileging diverse contributions, as well as focusing on themes that go beyond communication processes and museum spaces.

The museum is a place where: "sensations, ideas and ready-made images radiated by objects and gathered references together illuminate essential values for the human being. A fascinating space where one discovers and learns, in it one widens knowledge and deepens awareness of identity, solidarity and sharing" (Santos 2014, online, our translation). Wagensberg (2005) proposes a definition conceptualized as Total Museology; for the author, the museum is a tool for individual change and therefore for social change. Reality is built from both objects and phenomena, and the role of the museum is to stimulate.

The museum object is an actor and also an information agent. Through the stories told by these objects, memories arise for visitors, often from a past time. It is from contemplation or immersion that the visitor constructs narrative discourses. We can say that there is a museological temporality in the exhibitions. The museum objects are coated with the reality, meanings and cultural manifestations of man and it is in this perspective that the visitor builds networks of intentionality, through memories and recollections. In this sense, the perceptual act arises from the direct contact with the memory and from the encounter relationship of the subject with the object. In this scenario, understanding the visitors and their expectations is imperative to organizing the communicative strategies properly.

**Museum communication**

The discussions of new museological perspectives have also provoked the development of new approaches in the literature; not only as a result of the discussion of existing museum practices, but several attempts to renew or develop a brand-new museum frame were also presented (Alsford 1991; Bearman 1995; Anderson 1997; Scweibenz 1998; Hodge 2011; Campos et al., 2011). By considering museological digital environments, conceptualizing some terms in the context of museums is deemed necessary.

The term communication, according to Lasswell's general model (1948), is understood as the action of transmitting information between one or more emitters (E) and one or more receivers (R), through a channel (C). This conception has a wide range
of applications, encompassing not only human processes, but also communication processes across machines. The model proposed by Lasswell (1948) may be a reciprocal phenomenon (E→C→R) with interactive characteristics, or even a unilateral phenomenon (E→C→R), and in this way communication operates in time and not only in space, named transmission (Debray 2000).

The function of museum communication includes a wide range of activities (Mensch 1992), of which the main ones include exhibitions, scientific publications, educational programs and events. The "Interactive Experience Model", based on the perspective of the visitor, was proposed by Falk and Dierking (1992) and is based on interaction within three contexts: personal context, social context and physical context. The personal context embodies a variety of experiences, including visitor knowledge, which includes individual interests, motivations, and concerns. The perspective of each visitor is strongly influenced by the social context; individuals visit the museums in group, or alone, and invariably, get in contact with other visitors and museum staff. The physical context includes the architecture and "feeling" of the building as well as the exhibitions contained in the physical environment. In the proposed model, all three contexts are considered.

ICOM (2008) applies the term "communication" in a more complex way and includes the exhibition, publication and education functions of the museum in the museum communication process: "the museum […] acquires, conserves, researches, communicates and exhibits the tangible and intangible heritage of humanity and its environment for the purposes of education, study and enjoyment." This logic presents a more general museological communication process, observed in the PRC (Preservation - Research - Communication) process proposed by Reinward Academi from Amsterdam, which includes the functions of exhibition, publication and education exercised by the museum in the process.

We realize that research has advanced to understand communication in museums through user research and visitation studies. Usually the research responds to how the individual “uses” the museum, and although they bring important issues, these studies point out to perspectives limited to certain contexts.

This new way of thinking about museum communication and studies with visitors also meant a breakthrough with some traditional approaches to museums, therefore provoking a conflict between old and new generations of museum scholars (Candlin 2012; Kidd 2011; Johnson 2005; Marwick 1995). Perhaps because new approaches often involve interdisciplinary work and consultation of experts from other areas of knowledge, as Janes (2013, p. 15) points out:

The reluctance to use consultants and knowledgeable outsiders to tap different, and perhaps contradictory, perspectives to stimulate internal rethinking may be the result of professional conceit or insecurity among senior staff, but the consequences are the same - the loss of untapped internal
knowledge and creativity with which to build institutional renewal.

It is important to consider other communication means, in addition to the exhibitions, presented as the most important means and communication process for museums. Therefore, exploring new ways of thinking through interdisciplinary research and the use of communicational creativity in museums is desired. One of the tasks of the contemporary museum is to provide different types of learning for different audiences that can provide meaning, entertainment and possible knowledge construction through interactivity.

As new technology is introduced, new artifacts are created, interaction through digital environments develops rapidly, and many museums are embracing these new media as part of their communication (Tischritzis and Gibbs 1991; Hornecker and Stifter 2006; Tsitoura 2010; Campos, et al. 2011; Hodge 2011). There are relatively few studies on the ways in which individuals interact with exhibitions in museums and galleries, and even fewer studies on the form of collaboration and participation of subjects interacting with new forms and interactive media in museums.

In this way, some of the new perspectives that the museums have been facing involve the new media. Below, we present a structure that proposes to identify the problems and solutions for the development of knowledge organization systems in museological entities through Design Science Research.

**Design Science Research**

Design Science Research was introduced by R. Buckminster Fuller in 1957. The author defines research in design as a systematic way of planning and designing, whether for medicine, architecture, product design, graphic design, information systems, etc. The term was later used in Gregory’s book (1965) at the Conference on Methods in Design, when the author distinguished between scientific method and design method. Herbert Simon (1968) in his lectures on "Karl Taylor Compton" used and popularized these terms in his argument for the scientific study of artificial information systems.

Contemporaneously, a design research framework developed by Hevner et al. (2004) combines two paradigms of research: the paradigm of behavioral science and the paradigm of design science. Behavior science develops and verifies the needs of the digital system/environment, whereas design science produces an efficient and effective technological utility to solve the problem (Hevner, et al. 2004; Hevner 2007). The constructive activity in Design Science Research can be understood as human cognitive (e.g., complexity, creativity, control) and social (collaboration) activities, which contribute to the design and development of mediating objects, with the possibility of enhancing knowledge and information organization of information systems according to the context and human needs (Hevner, et al. 2014).
The framework and the conceptual guidelines used in the present proposals are presented in Figure 1 and 2, respectively.

Figure 1: Framework for Design Research in Information Systems (IS)

Simon (1996) points out that Design Science Research is motivated by the introduction of new artifacts and the processes that allow construction and innovation from these artifacts. According to the research developed by Hevner (2007), three research cycles are identified.

To understand the framework, the definitions and meanings of each of the cycles is presented. The environment defines the space of the problem (Simon, 1996) where the phenomena of interest reside, in addition to defining the business needs, it is where the relevance cycle is found. The knowledge base provides the raw materials through which research in information systems is performed. The knowledge base is composed of scientific foundations and methodologies for investigating the problem. The methodologies provide guidelines used in the justification/assessment phase of the problem situation. In behavioral science, the methodologies are typically rooted in data collection, empirical analysis techniques, case studies, and experimental simulations.

According to Hevner (2007, p. 3):

the relevance cycle initiates design science research with an application context that not only provides the requirements for the research (e.g., the opportunity/problem to be addressed) as inputs but also defines acceptance criteria for the ultimate evaluation of the research results. Does the design artifact improve the environment and how can this improvement be measured?
The result of scientific design research should be applied in the environment specific to the context which is found for study and assessment in the application and/or information system. The study field of the artifact can be carried out through appropriate methods of technology transfer assessment, such as action research (Jarvinen 2007).

Rigor Cycle connects the Design Science activities with the knowledge base of scientific foundations, experience and expertise that bring information about the entire research project and the state of the art in the context of the research object. Through this knowledge base, specific knowledge is provided. This should ensure the development of an innovation in the field of knowledge the research is contributing.

Design Cycle, located in the center of the framework, iterates and interacts across the main activities of constructing and assessing the artifacts and across the processes that involve the research project of the artifact in question. The three cycles should be present and clearly defined in a research project in Design Science.

Design Science is inherently a problem-solving process. One of the fundamental principles of Design Science Research is based on the knowledge and understanding of a Design problem and its solution, the guidelines developed by Hevner, et al. (2004) were based on these principles and are described in Figure 2.

According to Hevner, et al. (2004) the guidelines are meant to assist stakeholders to understand what an effective design-science research comprises. The authors are against a strict use of guidelines, as they believe researchers, reviewers and editors should use creative skills and personal judgement in determining the use of each guideline in a research project.

![Figure 2: Guidelines for Design Science Research](Source: HEVNER, et al. 2004, p. 83)
The Framework and the guidelines (Figure 1 and 2) provide the specificities to be included in the steps and subsequent definition of the design metrics to solve the mentioned problems. Thus, the guidelines should inform the development of an efficient and effective information system according to the human needs found in order to optimize knowledge organization.

Conclusions

The postmodern museum is clearly a participatory and attractive museum (Black 2005; 2012; Simon 2010). Thus, the museum needs to undergo transformations and have some flexibility. Many museums seek to be collaborative and interactive and, at the same time, they must be places that provide social interaction, the construction of relevant meaning, and places of experimentation and curiosity. All these elements cover a flexible form of communication. In order to construct a strong and robust new museum communication, museums should implement a thorough communication strategy. The transformations in museums have been underway for some time, although museums are always different and approach new features in their own way. Most museums could benefit from articulating their goals and purposes in redefining communication and using the principles of Design Science Research.

Hornecker and Sifter (2006, p. 7) highlight that:

One of the proponents of ‘Experience Design’, Nathan Shedroff (2000) argued that the most engaging interactive experiences allow for productivity, creativity and/or communication, as these (a) are basic human motivations, and (b) inherently entail interaction by requiring open-ended activity of users. This leads us to suggest creative, communicative and personal interactions as a valuable avenue for installations in public spaces (museum installations seem a useful test bed for exploring these types of interactive experiences).

In general, the organizational structure of information appropriation in different contexts and media in digital environments of museological entities demands a re-contextualization of the museum object when it is transformed from the physical to the digital context. In this process, the interface is the locus where the interaction of the individual with the museum objects and the knowledge organization content of this digital environment take place. This interface should, therefore, be based on the principles of Design Science Research framework and guidelines to provide quality solutions and innovations, pointing out new perspectives for communication and mediation of museum objects.

References


Tiago Trindade Cruz

Digital heritage: challenges and opportunities in the access and organisation of digital knowledge in contemporary societies¹

Abstract
This presentation aims at expounding on the importance of information and communication technology (ICT) and of digital heritage in strengthening the bilateral relations of and the interaction between civil society and its collective cultural heritage, as well as their role in the recognition and understanding of the particularities and the importance of the history and accounts of these constructed sets. By following this approach, I intend to address the subject of digital technology, highlighting its ever-increasing importance and leading role as an instrument for the study and preservation of cultural and architectural heritage assets. The latter aspect is particularly vital in understanding cities as living entities and physical embodiments of the collective memory, so as to create a database and a set of various highly significant types of knowledge, which can potentially be universally accessed and shared.

The use of ICT, providing new forms of accessing and organizing knowledge, has become an increasingly central and essential process in the field of digital heritage. The management of content disseminated through ICT acts as a basic element in the overall communication and sustainability of knowledge. ICT not only broadens the scope of research but also serves to disseminate it to a wider and more diverse audience in an interactive manner.

On the other hand, the implementation of research projects in the field of digital heritage – based on history of art and history of architecture – leads to the exploration of approaches with the aim of recovering, analysing and interpreting lost or otherwise invisible/transformed heritage assets within the urban landscape. By extension, they allow us to not only question how these strategies are used to reinsert the absent/transformed historical city in its multiple layers within the contemporary environment but also to assess the connection of civil society with its (in)visible legacy. In conclusion, the multidisciplinary nature of this approach is the key focus, an approach that effectively combines digital humanities, history, history of art, architecture and information technology.

1. Introduction
Contemporary society, in the context of the general improvement of its competences in the digital environment, tends to show great predisposition (the greatest in its history up to now) for sharing and preserving information – a factor that contributes immensely to the growing public interest in this subject (Thurston 2013, 31). In the specific field of cultural heritage, such an inclination raises multiple issues concerning the manner in which individuals may be keenly involved in exploring and promoting this area, as well as in the organisation and diffusion of knowledge per se.

¹ This paper is part of an ongoing PhD. research project in Heritage Studies, with a specialization in History of Art, at the Faculty of Arts of the University of Porto. The aforementioned paper is being developed under the supervision of Maria Leonor Botelho (FLUP - DCTP/CITCEM), and is co-supervised by Teresa Cunha Ferreira (FAUP – CEAU). The project is currently being financially supported by FCT, and has the following reference: SFRH/BD/132302/2017.
Since cultural heritage is a highly interdisciplinary field, this paper will combine and establish relations between the fields of digital humanities, history, history of art, architecture and information technology, laying the groundwork for cooperation and the development of strategies shared between the various fields involved. I will begin with a general outline (including a brief literature review) on the subject of research and methodology in the field of digital heritage, followed by an interpretation and acknowledgement of the challenges and opportunities that digital technologies present, specifying their applications in digital heritage.

Finally, I will conclude this paper by presenting a set of various concrete hypotheses for the development of knowledge organisation, with regard to its practical applications. As a final note, one that will evoke the general line of thought present throughout this paper, I will highlight the impact of digital technology and its place in contemporaneity in a context that is simultaneously cultural and social, under constant development, promoting discussion of its nature and ongoing redefinitions.

2. Challenges and opportunities of cultural heritage in the digital age

2.1. Digital Heritage: General Principles, Object and Methodology

Concerning Digital Heritage², the use of information and communication technology (ICT), framed within new means of accessing and organising knowledge, has assumed an increasingly central role, not only as an instrument for knowledge and dissemination of cultural heritage and, more specifically, in the manmade heritage assets of cities (a prominent example of this phenomenon) and the historical continuum, but also as a tool that allows innovation in society’s relationship with culture, in particular, for those who are keenly interested in cultural, material and intangible heritage³. With its ability to deliver beneficial and meaningful experiences, the use of ICT enables one to take part in rewarding – and potentially more integrative – activities, involving the five senses and allowing for an understanding of the historical past from a sensorial, dynamic and interactive perspective. Thus, the management of content disseminated through ICT functions as a basic element in the overall communication and sustainability of knowledge; it broadens the field of research and contributes to its dissemination to a wider and more diverse audience in an interactive way.

From a methodological standpoint, one of the most relevant issues is the concrete and specific content that the field of digital information could offer to research itself and its digital preservation. In digital heritage specifically, the creation of digital models provides opportunities for testing of information collected from documentary,

² “Digital heritage is made up of computer-based materials of enduring value that should be kept for future generations. Digital heritage emanates from different communities, industries, sectors and regions.” (UNESCO, 2013). See “Charter on the Preservation of Digital Heritage” (UNESCO).
³ Please see the paper titled “The Intangible and the Digital” (Hennessy 2013).
iconographic and archaeological sources, in a virtual sphere that enables us to recreate facts such as urban implantation, scale, layout and the interior and exterior design of buildings, as well as the environmental, spatial and landscape elements that integrate and/or surround them. In digital platforms it is possible to propose a recreation, to discuss it and to update it accordingly, all at a very low cost. Hence, the mastery of digital tools serves to further facilitate how cultural knowledge can be scientifically updated and channelled into didactic and recreative spheres as well as being disseminated in research projects concerning the field of heritage.

Consequently, this approach demands an effective understanding of the values and meanings conveyed by architectural heritage and achieved through historical knowledge – according to the International Charters and Conventions (ICOMOS, Council of Europe, UNESCO), in their readings and interpretations of the material expressions of human action. Based on the polysemic meanings of cities and their cultural heritage, digital reconstitution processes constitute feasible, non-intrusive, versatile and completely reversible solutions for processing knowledge of manmade heritage, in diachrony and synchrony. Moreover, such interpretation brings forth a certain degree of novelty in that a holistic reading of an object of study conveys a precise and framed description of knowledge and attempts to blur the existing distinctions between scientific research and how it is interpreted and presented to the public.

In this digital research approach, I highlight the use of different tools of analysis and interpretation, using drawing as a research tool (geometric, constructive, stratigraphic, among others), providing a basis for registration and three-dimensional reconstruction. Thus, the development of research in this field enables the public and researchers alike to access to open source information content through the internet or other digital means. Digital research permits us to overcome difficulties regarding the impossibility of access to the sources and sites studied, as well as the recording of previous phases and the current state of conservation of manmade heritage assets.

2.2. Digital Heritage: challenges and opportunities

As I have previously stated, in contemporary society we face a multitude of challenges when attempting to access and organise digital knowledge – from the involvement of the public and the issue of semantic harmonisation to the ways in which it is stored. On the other hand, these challenges point us to the very origin of said knowledge, be it a text or a web page. One must keep in mind that, as the "Guidelines

---

4 See “Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years” (Thibodeau 2002).
5 See the project titled “Arquivo.pt”, a project dedicated to the safeguarding of files extracted from the web (since 1996). It also allows the general public to peruse its repository. See also: “The Importance of Web Archives for Humanities” (Gomes e Costa 2014).
for the Preservation of Digital Heritage" (UNESCO 2003) state, much of the knowledge produced and processed today exists exclusively in digital form, presenting new questions regarding its organisation and preservation. In addition, as the "Charter on the Preservation of the Digital Heritage" (UNESCO 2003) affirms, there are also the issues of both the recognition of digital information as a legacy for future generations and the awareness of the risk of loss associated with digital heritage.

One of the greatest challenges posed by digital preservation is the fight against the fragmentation of the activities involved and the dispersion of available resources (Ortelli et al. 2017, 7). The encouragement of dialogue and structured interactions between the various parties involved in the process is, therefore, of the utmost importance – as is establishing "a «praxis» that is increasingly based on cooperation and interdisciplinarity" (Pinto 2005, 53), one that promotes the profitability of resources. The investment in collaborations between researchers in the fields of human sciences and ICT specialists must be incentivised and strengthened, by appraising the individuals involved in significant advances in the field of digital information – particularly in the field of digital heritage. It is essential to maintain good practices in all stages of this process, aiding in the clarification of "places" where they should be carried out, on the basis of transparency and accountability (as is mentioned in the London Charter 2014). These good practices should extend to all digital domains.

Digitisation is more and more viewed as an opportunity rather than as an obstacle or an enemy; there is a certain degree of global awareness to its potential benefits (Thurston 2013, 31). At the same time, this new perspective must be accompanied by its agents, those of the previously mentioned culture of sharing, rewriting skills and profiles, instilling in them the spirit of co-creation (Ortelli et al. 2017, 7), in favour of the "emergence of the social information system - a single and indivisible entity" (Pinto 2005, 53).

Another vital aspect is information management (Ortelli et al. 2017, 7), because of data and model interoperability, and the issue of semantic harmonisation. One also faces severe challenges in the acquisition and application of information, due to its applicability to a transnational universe, with the need to reconcile different types of repositories, databases and tools. It is also important to consider the physical storage of digital data, which introduces additional issues related to technological evolution and the necessity to guarantee the preservation of information in conditions of integrity and

---

6 See the Centre for Digital Heritage. It is composed of scholars from various universities and these researchers come together to conduct interdisciplinary research in the field of Digital Heritage [https://www.york.ac.uk/digital-heritage/].

7 See the U.OpenLab – with a pilot project centred on the Digital Museum of UPorto – developed at UPorto, in the context of learning/teaching. It asserts itself as a platform of mediation between the needs of the real world and the context of the University’s practical work (Pinto et al. 2016).

8 See “The Diverse and Exploding Digital Universe” (Gantz 2008).
authenticity (Pinto 2005, 53).

Preserving the memory of damaged/destroyed cultural heritage, protecting it and preserving it for future generations⁹, is one of the greatest challenges faced by knowledge organisation in the field of digital heritage (Ortelli et al. 2017, 7). It is equally important that people themselves come to realise that memories are (or should be) collected. The challenge is, on the other hand, to understand how to collectively present the amassed data and how to disseminate it via the various channels available.

In essence, the implementation of research projects in the field of digital heritage – based on history of art and history of architecture – leads us to explore approaches with the aim of recovering, analysing and interpreting lost or invisible/transformed heritage assets within the urban landscape.¹⁰ This implies and presupposes the abovementioned interdisciplinary development shared with the local populations – who should play an active role in the whole process.

2.3. Digital Heritage: access and organization of knowledge

The digital preservation of elements for their use as cultural assets, capable of being managed and organised, faces several obstacles – for instance, the fact that the digital preservation of elements is not considered a development priority is one such difficulty (Thurston 2013, 32). On the other hand, the information-gathering process is highly significant, and the maximum quality of digitalisation must be assured, with a rigorous and effective quality control monitored by agents with technical and scientific expertise in this field. This domain requires professional skills to ensure the quality of the product, and all individuals involved must have specialised training in this field. Currently, archival specialists are also active members of the information community, with the capacity to develop a dedicated line of thought concerning new possibilities in terms of storage and archiving, as well as planning their responses to the challenges that the fast-paced and ever-changing technological development demands (Stancic et al. 2013, 108). In contrast, the current trend of digitalisation among archivists reiterates the need for preservation and gives rise to new focuses within the field of digital curation (Lee and Tibbo 2011; Yakel et al. 2011).

I have already discussed the benefits of open source availability of the information amassed by researchers for the scientific community. However, the issue of "open data" accentuates the lack of both standard ontologies and standardised vocabularies (Ortelli et al. 2017, 7). This particular matter clashes directly with the legal issues of

⁹ See “Visualizing Venice” (knowledge and communication of Venice, in digital maps and models) [http://www.visualizingvenice.org/visu/] and “City and spectacle: a vision of pre-earthquake Lisbon” (developed by the University of Evora, with the aim of recreating Lisbon as the city was before being ravaged by the 1775 earthquake) [https://lisbon-pre-1755-earthquake.org/].

¹⁰ See, for instance, the research work on the transformations that the conventual complexes in Porto were subjected to after the disbandment of the Religious Orders (1834).
international law. A highly relevant issue is that of copyright and intellectual property rights\textsuperscript{11}, one of the most tremendous challenges faced by digital preservation and access to information (Gard 2013, 46). Various studies show that this issue is itself a major challenge due to the relevant actors’ extremely limited knowledge of what it is indeed possible to do or not (Ortelli \textit{et al.} 2017, 7), with the agents involved experiencing doubts on multiple occasions.

On the other hand, the processes of preservation of digital information vary significantly according to the different types of objects to which they portray – text, numbers, image, video, sound, multimedia, digital reconstitution, among others. In the specific case of digitalisation, documents (such as maps or images) sometimes have non-standard dimensions, requiring therefore the use of specific – and extremely costly – scanners. Other issues are linked to factors such as the aforementioned copyright restrictions. Hence, it is imperative to devise technical solutions that mitigate these factors and allow for democratic access to these services of widespread interest to the population in its entirety, in the context of competing commercial interests. The involvement of organisations such as UNESCO – which has itself contributed to the digital preservation of indigenous societies, among others (Katikala \textit{et al.} 2013) – in the role of mediators in these issues is, therefore, of the utmost importance.

\textbf{Final considerations}

This paper had as its guiding principle the application of ICT to digital heritage and in it I attempted to explore its implications for the advancement of the field of knowledge organization. As I have demonstrated, researching and promoting cultural heritage in tandem with new technologies contributes to a better understanding of our common past, aiding its enhancement, preservation and protection. Accordingly, when cultural heritage is regarded as an aggregating element in the process of redefining the identity of communities and places, ICT can be a decisive factor in the ascertainment of the specificities and importance of history and accounts of its constructions and environments.

The widespread dissemination of knowledge, employing resources, methods and means conveyed by ICT, also allows cultural agents to be permanently aware of the need to preserve manmade heritage assets (or their memory) in the best possible conditions to access and appreciate their "forms and ideas"; that is, their materiality and multitude of meanings.

The scientific community has demonstrated its interest in these issues by promoting meetings and forums of researchers for the purpose of discussing methodologies and encouraging best practises. The last few years have been particularly fruitful to the fields of cultural heritage and digital heritage, with advances that have brought

\textsuperscript{11} See “Universal Copyright Convention” (UNESCO 1971).
digitization as a means of preserving the memory of the world to the centre of the discussion.

Preserving cultural heritage in the digital age requires far more than merely good solutions to the technical challenges presented by the digital information lifecycle (data capture, indexing, archiving and retrieval). The system of digitization and preservation is as valid as the investment made in the study and articulation between interdisciplinary research teams, covering the various fields of knowledge and assuring the quality of the preserved material. The main objective should always be to preserve the integrity of the information, allowing it to be viewed as both a whole and as a singular element. "It is intrinsic... to ensure that the information produced is authentic, reliable, complete, intelligible/usable and capable of being preserved" (Pinto 2013, 27).

From a historical perspective, the research and storage of information (and not merely in digital format) has focused on the record itself, defined as a document conceived or received in the course of a practical activity, and reserved for action or reference (Duranti and Thibodeau 2006). Indeed, when applied specifically within the field of digital heritage, this notion of records only serving as evidence of actions and transactions may be accompanied by a substantial loss of meaning in an isolated reading or one devoid of context. I believe that one of the greatest contributions of digital heritage for advancements in the field of knowledge organisation resides in the holistic reading of its object of study, seeking an integrated interpretation, contextualising all its various dimensions.

A review of the relevant literature made it possible to identify the scientific consensus regarding the notion that context is an integral part of the assessment of authenticity of archive records and the establishment of user confidence. The future in the field of digital heritage depends on the digital forensic knowledge of the past. Thus, a record must be understood as more than just a probative entity (devoid of meaning or read exclusively in its synthase). Semantics and meaning should play key roles in identifying synchronic timeframes and readings, defined not only by physical and sensorial factors, but also by the socio-historical dimension and by the spheres of individual and collective memory. At the same time, procedures (based on good practises and interdisciplinary cooperation strategies) should be implemented to establish principles of authenticity and reliability of evidence and digital records. More than the preservation of cultural heritage – the very preservation of culture and its manifestations are what is, indeed, at stake.
References


Enhancement of digital heritage through digital social networks

Abstract
This article questions the role of digital social networks in heritage mediation. It is based on a cross-analysis of semi-structured interviews and on the Facebook pages of two libraries (French National Library in Paris & Inter-university Health Library). Using a discursive-semiotic analysis and sociological interviews, it highlights how the use of digital social networks responds to a set of communication and mediation objectives while building a "mediation of proximity".

Since the mid-1990s, libraries have been digitizing their written and iconographic heritage and publishing it on their websites, portals or databases. Through online publications, library heritage is being transformed into images or texts that are easily reproducible, transmitted and reusable, thus continuing the process of “trivialization” of heritage in new ways (Jeanneret 2008). Since the advent of the social web in 2004, some libraries have begun to promote this digitized heritage on social networks. Initially developed for individual and entertainment use and as a communication device for utilitarian purposes by institutions and companies (Mlaiki et al. 2012), Facebook is the main social network chosen by libraries eager to increase their audience (Chevry and Slouma 2016).

In this paper, we will investigate the links between libraries and digital social networks based on the specific example of Facebook and its use by two libraries; firstly, the BnF (National Library of France in Paris), which today holds two Facebook accounts, one institutional and the other dedicated to its "Gallica" digital library, and secondly, the BIU Santé (Inter-university Health Library), which brings together three thematic clusters: history, medicine and pharmacy. This research is based on semi-structured interviews we conducted with Facebook account managers and discursive-semiotic analyzes of the three accounts between 01/09/2017 and 31/12/2017. The simultaneous use of these two approaches allows us to compare the communication and mediation ambitions of libraries with what is displayed on the screen, while answering a set of questions such as: How do libraries put digitized heritage into circulation through digital social networks (DNS)? What is the role of published images and texts on DNS? Does the choice of images and texts reflect a segmentation of audiences?

Animation strategies of the Gallica and BIU Santé communities
The analysis presented in this first part is based on two semi-structured interviews conducted with professionals in charge of digital communication on Gallica and BIU
Santé’s social networks. Gallica’s communication with its 350 French and foreign partners differs from the BnF’s institutional communication on social media. Nevertheless, they cooperate actively on reciprocal interaction days, during which a game is played across these two digital accounts, contributing to the impression of complicity between the BnF and its digital library and overall coherence. This communication is managed by a team of professionals across the various sites of the BnF, who in turn moderate the digital social networks and communicate with each other via chat.

Gallica’s communication through Facebook, Twitter, Pinterest and Instagram has two main objectives: first, to publicize and enhance Gallica’s 4 million digitized documents; second, to stay in close contact with the large community of ‘Gallicanauts’ (users of Gallica who actively participate in the dissemination of their documents on the Web). Very responsive, always in attendance and authentic, the community of Gallicanauts is a strength for Gallica and goes so far as to promote the mediation by spotting treasure or unexpected items in the digital library, thus highlighting digital heritage. Their work can be seen through seven videos entitled "Gallica and me: when the Gallicanauts talk about their digital library". In these videos, they share their rich and diverse uses of Gallica. These users propagate Gallica's heritage in blogs, personal, associative or institutional websites as well as on Youtube channels.

Gallica’s editorial line emphasizes links with its scanned documents. This mediation on digital social networks is a major tool for increasing attendance on Gallica, an objective determined by the tutelage of the Ministry of Culture and Communication. The tone of the publications is humorous and denotes a particular proximity with the Gallicanauts. High-quality statistical monitoring promotes a good knowledge of the public of each media. The public who follows Gallica on digital social networks is not necessarily the Gallicanauts. It is composed of French-speaking international internet users, 54% of whom are women, the majority being between 25 and 34 years old. Targets vary from one medium to another. On Facebook, there is no specifically defined target. On the contrary, on Twitter, it is more possible to reach people who do not know Gallica but are interested in these topics. On Instagram, opened in 2018, young people are targeted.

As for BIU Santé’s communication on digital social networks, it has four main objectives: first, to follow the movement by insisting that BIU Santé has contemporary documentation; second, to multiply the channels of communication and interaction with the library’s readers and to help a new public discover the library; third, to improve referencing and visibility; lastly, to share monitoring of (external and internal) information, especially on Twitter.

---

1 Interview n. 1 on 25/01/2018 with the web project manager of BIU Santé; interview n. 2 on 02/02/2018 with the general coordinator of Gallica’s digital mediation within the cooperation department.
Wishing to unite the clusters of medicine, pharmacy and history in a single establishment while balancing these three areas, it was decided in 2012 to create a single account on each social medium. The basis of digital social network communication is the blog, whose posts are subsequently rebroadcast on Facebook, Twitter, Google+, Periscope, LinkedIn and Viadeo. While Twitter functions as a relay for all blog posts (700 posts in 2012), Pinterest is the place of publication for all images picked-up and shared. The content of the information differs depending on the media used. The tone remains neutral, slightly eccentric, especially on Facebook and Twitter and without mention of the name. Images are used in an offbeat way and opinion statements are encouraged when they are sustained by arguments.

The user is brought into the picture through content that promotes interaction and commitment. For example, games are organized every year. The game "Pharmacies of the world", for instance, encourages the user to send the most beautiful pictures of pharmacies encountered during his travels. In this case, the target audience is individuals who are familiar with the library. Coordinated by the web project manager, a team of 15 people is involved in digital communication on different levels. Publications are grouped in a shared Excel file. At each publication, the information is written in the file. The project manager decides what should happen and at what time. While only two or three colleagues with administrator status write blog posts and moderate them, dozens of colleagues intervene in various ways. Some simply relay the information without producing content. There is no communication service.

In short, both Gallica and BIU Santé have a clearly defined social media policy. The choice of objectives, the editorial line and the target audience contribute to building the identity of the institution. This is the case of BIU Santé when it decides to present the three clusters in a unified manner by choosing a single account on each media platform. This is also the case with Gallica, which has decided to represent itself publicly with its partners.

According to the specialists interviewed, digital networking makes heritage more visible, it democratizes it and increases awareness among the public (interview n. 1). However, they report two difficulties. First of all, the institutions surveyed mention the change in the Facebook algorithm which has led to a drop-in visibility of posts. They feel heavily dependent on the changes instituted by the Facebook company (interview n. 2) and deplore the difficulty of having institutional pages visible on Facebook without payment (interview n. 1). In addition, many users share poor quality images from the BIU Santé on Pinterest without specifying their origin. In order to avoid this re-use of digital heritage out of context, which decreases its value, BIU Santé has chosen the open license Etalab². It allows the re-use of content provided its origin is

---

² As part of the government's policy to bring public information into the open ("Open Data"), Etalab has designed the "Open License". This license facilitates and encourages the re-use of public data made available free of charge.
The importance of the “text image” in the construction of heritage

A discursive-semiotic analysis of Facebook publications between September 1st and December 31st 2017 provides an overview of the strategies deployed by the two institutions. In this second part of the presentation of our research results, we return consecutively to the specificities of the digital communication of these two organizations. We will show that "text image" (Souchier 1998) plays an essential role in the manufacture of heritage.

Facebook at the BIU Santé: how daily communications naturalize the role of heritage as a communicative object

Although the institution does not publish on Facebook on a daily basis, this network is the subject of a significant investment with a multi-weekly publication (64 publications between 01/09/17 and 31/12/17). The "information" conveyed reveals a somewhat traditional typology for a place open to the public: it provides mostly news about the library (opening times, closures, new subscriptions, events) and promotes events organized by partners. In addition to these publications, the library's Facebook page is punctuated by the monthly publication of a calendar, featuring heritage objects. The publications always consist of a short text and one or more image(s) associated with a hypertext link either to a Facebook post or to the website of the BIU Santé. Within this framework, the homepage of the website contains a set of go-between signs (Candel and Gomez-Mejia 2017), embedded almost exclusively in written texts. They signify hypertext links leading to other pages of the site, principally to headings of the "search" section on the site biusante.parisdescartes.fr: "Our catalog", "Online journals", "Other Articles and Databases", "Electronic Books", "Theses", "Other Catalogs and Libraries"; or to sections of the site dedicated to "quick search": "Online Journals", "Medico-Dental Theses", "Medic @", "PubMed", "Sudoc". The mediation of heritage takes place through an archetypal and traditional set of forms and documentary and librarian formats.

On the Facebook page, heritage is often perceptible through the images that make up the publications, regardless of their informational objectives. For example, the promotion of European Heritage Days allows for the recognition and enhancement of old works exhibited at the BIU. In the same manner, the monthly calendar gives access to news from the site Medic @. The anniversary of the BIU also enables the institution to turn a painting of the first curator of the BIU into a “gobetween sign” linked to one

---

3 Association approved by the Ministry of National Education. It aims to promote the free sharing of knowledge.
4 Wikimedia Commons is a multilingual online media library launched in 2004, which aims to provide a central directory of free content (images, sounds, videos, audio, etc.) that can be easily used by all Wikimedia projects.
of his speeches, etc. On the other hand, on the BIU website, notices coexist with images to which they often give access. A tension is thus established between image and text, the nature of their relation depending on the type of media support: on Facebook, the image is the point of entry in terms of mediation. On the other hand, on the website, the image coexists with lists (notices), more readily favoring a classificatory logic.

All these publications, by being anchored in the daily life of the BIU Santé, enable it to highlight its collections and to put them in circulation. Several rationales are intertwined: a mediation logic (mediating contents by making them visible and accessible - for example by diverting them into a monthly calendar), a communication logic (institutional communication, also through links to other institutions such as BnF and its Gallica platform), and a technical logic (referrals to the site contribute to its natural referencing function).

Facebook at the BnF: between an institutional profile and a page dedicated to numerical heritage, a pedagogical and entertaining discourse on heritage.

The BnF invests greatly in digital social networks, especially Facebook. It has two pages: an institutional page "BnF - National Library of France" and a page dedicated to its digital library entitled "Gallica" (with 202 and 138 publications, respectively, between 01/09/17 and 31/12/17). The first is more consistent with the logic of promoting all events related to the life of the library such as opening hours or communication related to temporary exhibitions. The second, on the other hand, is entirely dedicated to the digital library and promotes and circulates the different collections. Cross-analysis of the two Facebook pages reveals similar ways of enhancing heritage. It is often mediated through posts articulating mediation (explanation of an element of heritage, presentation of a collection, etc.), images (of the mentioned heritage) and hypertext links. This triad is found specifically on the homepage of Gallica, and is especially relevant to understand how digital social networks can contribute to enhance libraries’ heritage. This triad unfolds around a set of information combinations:

<table>
<thead>
<tr>
<th>Text</th>
<th>Image</th>
<th>Hypertext link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catchy phrase such as: « Do you know (...)? »</td>
<td>Page(s) of a book, manuscript, etc.</td>
<td>Viewing the page on the Gallica platform</td>
</tr>
<tr>
<td>Presentation of a new collection</td>
<td>Images presenting the collection</td>
<td>Dedicated data bank (documentary portal)</td>
</tr>
<tr>
<td>Presentation of a type of object, from a collection</td>
<td>Image of medals, for example</td>
<td>Search page (search with set parameters) on Gallica</td>
</tr>
<tr>
<td>Mention of a current event, put in perspective with its preceding avatar (JO 2024 / JO 1924)</td>
<td>Images of this event in the past (no current pictures)</td>
<td>Search page (search with set parameters) on Gallica</td>
</tr>
</tbody>
</table>
In addition to these informational schemes, series can also be found. One of them is "the Friday Enigma" that invites the public to dive into the collection to seek "various facts from Gallica", which enables the institution to enhance digital heritage through the Gallica blog. The Gallica page is thus marked by several essential elements: an entertaining tone (both in the texts and in the proposal of a weekly puzzle punctuating the library’s publications); a reflection on temporality, through games on dates (for example, the date of December 14th inspires Gallica to project us into an old epic: "December 14, 1782, the date on which the Montgolfier brothers succeeded for the first time to fly a balloon inflated with hot air [...]”); an economy of information and mediation always referring to the Gallica platform and to the digitized object.

These two examples demonstrate how heritage is written (Tardy 2009) and continues to be written through hypertext links between media devices (website / digital social networks). Unlike the term “link”, which assumes the obvious function of establishing a relationship (Davallon and Jeanneret 2004), mediation strategies are deployed in different ways but display a set of invariables: the strong relation between image and text and the introduction of a publication rhythm, which is not limited to publishing regularly but also consists of creating opportunities for linking (a monthly calendar for the BIU Santé, daily puzzles for the BnF). The pace and extent of communication seems to be associated with libraries’ regulations, strategies and means. On both types of social networks, heritage exists through highlighting its "places" of storage. These are archetypal forms of knowledge mediation, whether documentary forms ("Image bank") or media forms specific to digital mediation (use of PDF, display of scanned page in a software, etc.).

Digital social networks reinvest and enhance forms belonging to the mediation of knowledge through software and structures specific to the digital realm. Thus, a process of "standardization" of heritage, through which it appears to merge into ordinary communication, is at stake in this use of social networks. Heritage becomes a digital object like any other, an artifact subject to the mass media economy in which visibility is crucial. If there is indeed a mediation logic specific to knowledge institutions, there is also a logic of media coverage. The communication of BIU Santé and BnF oscillate between two rationales: communication and mediation.

**Digital social networks: between a strategic investment and the renewal of a knowledge mediation ideal**

The "digital presence" of libraries denotes both a strategy of institutional communication – in which communication is intrinsically instrumentalized – and a desire for libraries to renew the way they enhance heritage, that is, an interest in mediation. This last section will explore these dynamics.

From a strategic point of view, the presence of libraries on digital social networks is an investment: it garners a team of dedicated people and indirectly mobilizes people
(15 people in total at BIU Santé) to respond to a directive. Indeed, it seems essential today for such organizations to be present on digital social networks. Libraries also choose to invest in them to meet another requirement: visibility. Considered here as an "instrument" of communication, Facebook allows libraries "to exist" in a more global context; a framework in which they can use social networks to engage in information monitoring, publicize this monitoring and thereby position themselves in a specific social, semiotic and technical framework.

The presence of libraries on social networks also responds to a set of communication, organization and mediation objectives. From an organizational perspective, BIU Santé encourages the unification of three main clusters: medicine, pharmacy and history. At the BnF, the Gallica page connects the library with a set of partners. Thus, these two pages allow two plural institutions to speak with a single voice, responding to the institutional imperative of incarnation (Rondot 2015). In this way, the Facebook page, besides being a strategic tool used to render the library visible, contributes to building the institution, giving it a voice. The institution exceeds the organizational limits of a cluster, a library or a university; it is the result of a process of networking of places sharing similar goals.

From a communication point of view, the two Facebook pages give visibility to the organization but also to its various communication devices: they open up libraries to digital social networks, and thus to their users, and simultaneously satisfy a referencing need. The two pages refer first and foremost to libraries’ other communication devices: their website and blog. These are privileged tools because they are considered as storage places for editorialized content and digitized objects.

From a mediation standpoint, these two Facebook pages enhance the patrimonial approach of the two institutions: digitized heritage is omnipresent on Gallica and referrals to catalogs or data banks are very common on the BIU Santé. Therefore, beyond providing information on places (whether digital or not), the two libraries’ Facebook pages are gateways to specialized content, thus operating a form of "popularization" by re-contextualizing heritage in an everyday space, more related to its users’ daily practices.

Hence, this re-contextualization of heritage on Facebook engages in the process of bringing digitized heritage into mediation. The use of the term "bringing into mediation" rather than mediation, is meant to highlight the strategic dimension of the investment of digital social networks by libraries: it is not only a question of following a ruling, but rather of proposing a genuine editorial policy consisting of increasing the value of heritage and mediating it on social networks. In this context, the use of Facebook, and social networks more generally, reflects a desire for renewal on the part of library institutions. Mobilizing community imagination and the networking capabilities offered by this network, they construct a "mediation of proximity"
composed of an entertaining tone, a "community of Gallicanauts" and a weekly “enigma” at the BnF, as well as a monthly calendar at BIU Santé. While this process is less pronounced at the BIU Santé, the BnF constitutes a particularly interesting case of deploying a global mediation policy on all these communication devices, at times almost becoming a transmedia logic.

References


Moisés Rockembach, Armando Malheiro da Silva

Epistemology and Ethics of big data

Abstract
Much is mentioned about big data, a trend term that defines large datasets that people have been producing with the use of technology. Indeed, there is an ongoing transformation in the processes of information generation and dissemination with this constant technological presence in our lives. As a research aim, this paper studies the possibilities of a relation between big data, epistemology and ethics, from bibliographic research. Information science, which has long been concerned with studies of the flow, organization and informational behavior, now coexists with data science, which has a statistical approach to the information collected, as well as the use of machine learning techniques and predictive analysis. In our study, it is not a question of distinguishing 'data' from 'information', but correlating areas that are increasingly approaching, based on technological convergence. Accordingly, some of the issues that most concern us in the use of big data are epistemic questions about data and their ethical uses, which deserve further studies.

Introduction
Epistemological reflections focus on determining the necessary and sufficient conditions for knowledge production and the investigation of knowledge sources for big data and its characteristics. Ethical reflections arise from use and manipulation of these large data sets and their implications, mainly related to privacy. Therefore, the fields of information science and data science intersect, seeking to investigate data-information sets on an increasing scale. The combination of these reflections aims to contribute to discussion of the scientific, construction of knowledge, and pedagogical field, bringing possibilities of studies and academic uses.

The paper has a qualitative approach, reflecting on the possibilities of relations between big data, epistemology and ethics, in order to demonstrate convergences between the object and the respective fields of knowledge.

The methodology of data collection adopted was bibliographic research, with keywords ‘big data’, ‘ethics’ and ‘epistemology’, based on a literature review from scientific journals and databases. The analysis and presentation of the results show relevant contributions of this bibliographic review, as well as the convergence and differences between data science and information science and the epistemological and ethical aspects of big data, emphasizing the need for these approaches to undergo better data analysis.

Big Data, Data Science and Information Science
Since the model of the DIKS pyramid, also known as Data-Information-Knowledge-Wisdom Hierarchy (Ackoff 1989), widely used in the areas of knowledge management, information science and management, to studies such as Zins (2007), which worked
with 130 definitions of data, information, and knowledge, there have been many conceptual discussions about these terms. It is recognized that this model, although widespread, has criticism about conceptual problems and in its proposal as a hierarchy, as cited, for example, by Frické (2009) and Capurro and Zins (2006).

Looking for a convergence of 'data' and 'information' concepts, the definition of information that we used is a “structured set of mental and codified representations (significant symbols), created in a specific social context and capable of being recorded on any medium (paper, film, magnetic tape, compact disc, etc.) and, therefore, communicated in an asynchronous and multidirected way” (Silva and Ribeiro, 2002). This concept includes the concept of data, since the data is not only a raw quantitative information, but also has meaning. Thus, we have two concepts for 'data', one that is "the conventional representation, through coding, of information in order to allow electronic processing", therefore, no different from the conception of information, and another which "means the physical or electromagnetic impulse or vibration, seismic, etc. which through specific technological devices are converted into graphical representations (information)", in this way, distinct from the meaning of information (Silva 2006 and Rockembach 2017).

Information science, as an area that studies information properties, users’ behavior, flow and means of processing, aiming at access and use (Borko 1968), has a contemporary application, so that the research and performance of this field is constantly being updated by the technological innovations developed and made available. Song and Zhu (2017) propose to Ischools, a consortium of Information Schools dedicated to advancing the information field, three focuses on the study of data science, from user-based, tool-based, and application-based approaches, thus differentiating from the usually used approaches in Computer Science Schools and Business Schools.

Big data is an object of research with fundamental importance in the emerging area of data science, a field of knowledge that mixes computer science techniques, such as programming languages, use of distributed computing platforms, application of scientific methodology and mathematical and statistical models for data analysis. Three fundamental characteristics that stand out in these large data sets are the 3 v's, velocity, volume and variety (Zikopoulos et. al. 2011).

The big data lifecycle, as demonstrated by Jagadish et al. (2014), includes: I. Data Acquisition; II. Information extraction and Cleaning; III. Data Integration, Aggregation and Representation; IV. Modeling and Analysis; V. Interpretation. This scheme can be visualized below:
The data acquisition involves activities of data collection, in addition to checking the qualities of these data, as to their reliability and storage form in the database. In information extraction and cleaning, the data type’s heterogeneity and error checking in the data collected is verified, which involves preparing the data for an adequate analysis. Data integration, aggregation and representation involve the analysis of heterogeneous data, aiming at uniformity and convergence for results that aid in the resolution of a problem or interpretation of a specific phenomenon or situation. In the data modeling and analysis, the use of statistics for large datasets, involving characteristics such as data heterogeneity, the scale of the volume to be analyzed, production dynamics and data analysis are often taken into account in real time, along with the interrelationship between variables. Finally, the interpretation involves not only the presentation of analysis results, but also the role of the decision maker and the generation of new insights from the conclusions.

The need for an epistemological and ethical approach to Big Data

In the bibliographic research in the Web of Science database (WoS), by topic (TS), from two sets of terms "Big Data AND Epistemology"; "Big Data AND Ethics", there were 64 results for the first search and 266 results for the second search. We highlighted the main results by index of most cited and filtering off-topic texts. This can be
indicative of research trends and the most relevant issues that the scientific community has been reflecting on.

Table 1: Big Data, Epistemology, Ethics - papers most cited in WoS

<table>
<thead>
<tr>
<th>Big Data AND Epistemology</th>
<th>Big Data AND Ethics</th>
<th>Number of citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mittelstadt, B. D., &amp; Floridi, L. (2016). The ethics of big data: Current and foreseeable issues in biomedical contexts.</td>
<td></td>
<td>32</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Big Data AND Epistemology</th>
<th>Number of citations</th>
<th>Big Data AND Ethics</th>
<th>Number of citations</th>
</tr>
</thead>
</table>

The ten most cited papers of each research were listed, four papers were found in the
two searches among the most cited and appear at the beginning of the table.

The large volume of data that we call big data, according to Boyd and Crawford (2012), is more related to the new possibilities of research, analysis and cross-references than a volume of data that before the creation of this term did not exist, for example, census data. Therefore, still according Boyd and Crawford (2012), big data can be considered a cultural, technological and academic phenomenon, which is based on data processing technology, analysis for pattern identification and the mythology that large datasets provide intelligence and new insights, with aura of objectivity, truth and precision.

According to Abbagnano (2007), Epistemology, or Theory of Knowledge, as a field of philosophical knowledge, is approached and updated with the methodology, since it analyzes research procedures, validity conditions and limits, and the linguistic instruments of scientific knowledge.

In the epistemological approach, we need to understand what we mean by data and information. As previously stated, concerning the production and apprehension of content, we understand that it is always given a meaning, since the semantics is attributed by the subject, therefore we treat the concepts of data and information as the same concept, corroborating Silva and Ribeiro (2002) and the understanding of semantic information from Floridi (2010).

Another possibility of an epistemological approach that is demonstrated in the study and analysis of big data is the very understanding of epistemology as a philosophical branch that deals with human knowledge and is related to the philosophy of science. If the analysis of these large datasets is done by data science, which presupposes the use of scientific methodology, this also leads to thinking about the relations between belief and truth in science and how the diversity of data and its complexity affect the conclusions, also situating big data discourse within a broader epistemic and historical narrative (Rieder, Simon 2016). The use of classical epistemology methods now joins with machine learning and predictive methods of data analysis.

However, decontextualizing the data, which can result from quantification, classification and algorithmic analysis, can be problematic for data interpretation (Mittelstadt and Floridi 2016), since this loss of context may also mean loss of knowledge, which is attributed precisely by the contextual data. Crawford, Gray, Miltner (2014) argue that we should invest in an explicit epistemological pluralism and reflect on how data is generated, collected, and how this subsequently affects data analysis.

Regarding the knowledge sources for big data analysis, we emphasize the crucial challenge that involves data heterogeneity. Therefore, with regard to heterogeneity issues, the standardization of data to be processed to, in the end, provide new insights, is a major challenge for data scientists. The production of adequate metadata, the
structuring of the database and adequate metrics for statistical treatment, considering
different sources of information, from numerical data to digitized texts and images, and
the crossing of these data, are factors that demand solutions according to the problem
to be solved. One of the V's that characterize big data, that of 'variety', is directly related
to this research challenge, since it seeks to converge structured and unstructured data
for the same analysis and production of a new set of knowledge.

In big data research, according to Floridi (2012), the possibility of small patterns
analysis is an epistemological and ethical problem. In the big data ethical approach, the
use of this data can influence behaviors from the analysis of patterns (Floridi 2012).
The dilemma that is evident is related to data privacy and the need to protect personal
information, an ethical-legal feature reflected in many countries, but still with
regulatory gaps.

For Mittelstadt and Floridi (2016), five ethical risks appear regarding big data:
informed consent, privacy (including anonymization and data protection), ownership,
epistemology and objectivity, and big data divides, the latter related to "inequalities
between data subjects providing the material for big data analytics and the
Organizations with the necessary infrastructure and resources to analyze and
understand the data" (Mittelstadt and Floridi 2016).

As for privacy in big data, the confidentiality of large medical data sets, both in
databases and those produced by wearable devices, connected to the Internet of Things
(IoT), are ethical factors that are increasingly worrying. Other issues that deserve
mention are the reduction of personal data privacy, the possibility of a decrease in civil
liberties and an increase in state and corporate control, based on the manipulation and
use of large data sets.

**Concluding remarks**

Big data is used in a number of different areas, as we could verify in the many articles
found. This paper emphasizes important points such as the convergence between
information science and data science to study big data, the results of bibliographic
research and some of the main issues about epistemology and ethics of big data.

The convergence between information science and data science is visible, although
it does not merge into a single field of study. Data mining is a precise, well-calibrated
strategy of extracting data and converting it into a reading that has relevant statistical
bias. On the other hand, data science, which has grown and become more generalized
nowadays, has a restricted dimension that makes it dependent on the statistical
treatment of huge volumes of information. In this sense, data science cannot be
confused with information science, because the dividing line is clear. Since data are
information, there is a potential approximation between the areas. In conceptual terms,
data science has as its focus or object of study "information" but, methodologically, it
is not an applied social science, it is an applied technology. Information science cannot fail to include data (Silva 2006) in the broader concept of information (Silva and Ribeiro 2002) and therefore, by doing this, can expand its scope and its performance as applied social science.

Finally, questions about epistemology and ethics serve to establish a critical sense of the knowledge produced by big data analysis and limits on the collection and use of personal data. Epistemologically, one of the main issues is to evaluate how knowledge is being constituted from big data, taking into consideration elements like context, veracity, and interpretation, and with critical thinking about how information is collected, analyzed and reinterpreted. Ethically, the main question is how big data tools should be used and the implications for misuse of these tools by many actors involved in the process, from data providers to data analysis. The increasing use of big data as a means for analysis in diverse contexts deserves further studies, with ethical and epistemological contributions.

References


Maria José Vicentini Jorente

Information Design as knowledge technology in the organization of digital information environments on the Zika Virus and its effects

Abstract
This paper studies behavioral changes in the processes of knowledge organization applied to digital e-Health information environments on the Zika Virus and its effects in Brazil, from an interdisciplinary perspective combining information science (IS) and information design (ID). The objectives of the study are to review the literature related to knowledge organization in Web 2.0 e-Health environments; as well as to verify the occurrences of Web 2.0 platforms for on-screen purposes; to undertake netnography in WhatsApp environments related to the subject; to create controlled vocabularies from the verified folksonomy, based on the netnographic immersion; to study elements to translate complex knowledge; to create model recommendations for the knowledge organization of a digital environment on the Zika virus and Microcephaly. In terms of methodology, following the theoretical analysis, we carried out a netnographic study, through authorized immersion in a closed group of mothers of children with microcephaly in WhatsApp and extracted the folksonomy for the creation of controlled vocabularies. We then categorized the terms and proceeded, through colon classification, to make a faceted classification. The categorization provided a more comprehensive set of concepts and their organization in classes and sub-classes, essential for information analysis, synthesis and representation, providing the basis for information search and retrieval in an information environment. Four categories of subjects shared by the community of interest were extracted: a) medicine; b) treatments; c) clinical status; d) benefits from government / public institutions. The research is justified by the need for academic-scientific action in organizing knowledge within digital e-Health environments. The association between ID, IS and computing makes a field of interrelationships from which emerge new hybrid information elements, language and complex media convergences, and even new knowledge domains. Different approaches and paradigms coexist in the various domains to organize knowledge, structure information and guide searches, in environments that integrate the form, context, content and subjects of information, in social Web (Web2.0) environments.

Introduction
This paper is part of a research project carried out at São Paulo State University (UNESP), and involves the perception of behavioral change of the various social actors in the processes of knowledge organization in responsive, open, customizable and interoperable digital environments. From an interdisciplinary perspective between information science (IS) and information design (ID), this ongoing research aims to reflect on the problems of knowledge organization and its communication on digital platforms aimed at netizens in general and, therefore, to foster access to targeted knowledge in digital virtual environments. This research is developed within the scope of IS, which, similarly to ID, focuses on human needs. Besides its epistemological nature, we consider that, given the social scientific status of IS, the creation of a set of recommendations based on ID disciplinary resources would also be a desirable result. Thus, upon the creation of the project, due to the emergence of cases of microcephaly resulting from the epidemic caused by the Zika virus in Brazil, we directed one of its
specific objectives to the assistance of those seeking for information in the pediatric health area, with the aim of acquiring knowledge related to the consequent living conditions of children born to women who were at some stage of pregnancy when contaminated. Since this is a study in progress, the chosen methodologies and strategies are currently divided into two groups: 1) those that have already been implemented (with results obtained), and 2) those in the process of being implemented.

1. Theoretical production regarding Web 2.0 e-health environments

According to the methodologies and strategies already implemented, we have reached the following results:

a) The literature review on microcephaly from Zika, though not exhaustive, provided a first contact with the subject. We highlight a bibliometric study on the subject of the Zika virus in five international databases (Lilacs, PubMed / Medline, Scielo, Scopus and Web of Science). The results show a broad time frame, from January 1945 to January 2016. Their findings indicate a total of 336 publications, of which only 242 were relevant to our research. The scarce production on the subject was also verified when compared to other viral disease emergencies, in addition to the lack of Brazilian studies on the subject, demonstrating the need to develop national research directed to the confrontation of the epidemic associated with the Zika virus and associated conditions.

b) As for the formal processes of publication and sharing of targeted knowledge about the theme, we may reflect that, by externalizing knowledge, we represent our experiences from audiovisual resources, creating interactions and communications that influence the way cognition and thought react to each of the situations experienced and formalized during the sharing process. The efficiency of this process depends, therefore, on the formatting of representation languages, consisting of syntax and semantics, translated through information design (ID) into a design, or representation.

As for knowledge intermediated by the environments which emerged from the Web 2.0 transformation at the beginning of the 2000s, it has provided netizens with numerous interactive possibilities through a new typography of computational interfaces. The consequent response of netizens to the perceived need for efficient and effective design in the curation of digital formats and contents of web-accessed environments is to demand that knowledge be organized and translated through well-communicated, efficient information systems that can genuinely help people to accomplish tasks and acquire knowledge via the Internet. On the other hand, websites with knowledge and information organization modes based on traditional top-down subject-based tools, which may fail to meet the needs and demands of the information society, are still disseminating information, despite new possibilities of data curation through the systemic convergence of disciplines related to information representation.
This paradox is currently present on the Internet and is perceivable in Web 2.0 platforms. In general, the hypotheses for its solution comprise interdisciplinarity and convergence across areas of systems knowledge involved in information sharing and in building the contemporary liquid society, in which knowledge of information and organization is assumed. Among these areas are information science, knowledge organization and, also, especially with respect to issues of representation, information design.

From the Latin roots *designare* and *signum*, meaning developing, conceiving, the term *design* has a sense of project and the same etymological root as “desire” and “designate”. In 18th century, in England, the term re-emerged as a translation of the Italian *disegno* to denominate activities related to the production of objects in the Industrial Revolution. Since then, design has become an umbrella discipline, provided with methodological resources to achieve certain productive objectives. In the 18th century, graphical synthesis emerged and was developed as an effective means of organizing information, which became commonly used, while in the 19th century, the Nightingale polar graph and Minard's *Carté Figurative* introduced new ways of visualizing information.

At the beginning of the 20th century, Bauhaus spread a new design concept focused mainly on the relations between function and functionality and on the internationalization of an essentialist style. As a result, in the 1940s, Ladislav Sutnar, consolidated knowledge he had gathered to develop information design (ID) as a discipline. In the 1960s, Jacques Bertin explored intersemiotics, a theory of signs and other aspects of representation. From the 1980s onwards, with the massive rise in the creation of computational documents, researchers in ID have endeavored to study and understand the new technology of information dissemination, the emergence of skills required by such technology, the power it gives to the area of ID and its technical and functional specifics. Up until the 1990s, new perspectives were opened by usability studies, emerging from ergonomics, which had predominated since the 1970s. The focus of this perspective was to involve the subjects that deal with information and its formatting, its design in the new context of technological devices; that is, participatory design. In 1999, Horn argued that ID sought to organize information for its effective and efficient role in physical and digital environments, with the objectives of promoting understanding, accuracy and retrieval of documents by facilitating their transformation into effective actions and adding the value of interactivity to information retrieval processes. The relationship between information organization and retrieval and the design of interfaces and interactions was, therefore, narrowed.

For Tramullas, this was "the theoretical structure in charge of creating systems and interactive communication products of information oriented to the user, based on the design of digital documents" (2003, p.109-110, our translation). Bonsiepe (2011) and
Shedroff (2000), at the beginning of the 21st century, presented the ID discipline as a domain. In this domain, certain disciplinary resources allow the execution of representation tasks, aimed at data organization and presentation - through selection, ordering, hierarchization, connections and visual distinctions, with meaning and value and structure - so that information expresses its potential in an efficient manner and provides the means for individuals to collect and process it. In a complementary way, Garret (2002) created the classic scheme that placed the role of ID among the elements of user experience (UX), and identified parallel forces or areas in Internet systems, more specifically in the Web: software interface, task-oriented (ID). Garret further defined interaction design (InD) as a sub-area of ID, with regard to actions leading to certain human reactions to applications, to complete a task and achieve a goal. Thus, InD instructs the netizen’s actions and subsequent reactions in the accomplishment of the tasks necessary. The types of interaction are, for example, ringtones on the screen, audio help, voice activations, etc. Thus, InD helps the netizen to access the flow of information, via interface dynamics and functionalities. In this way, from the end of the 20th century, ID was consolidated as an autonomous discipline, with a body of specialized knowledge, being constituted as a sub-area of design (Burdek 2006). At the same time, Burdek (2006), Cross (1982, 1999, 2007) and Bonsiepe (2011) considered ID as a distinct discipline with a body of knowledge, content and research base of its own.

The development of ID was therefore based on the idea that the effectiveness of information communication for knowledge creation is related to an appropriate organization of its form. There have been multiple solutions provided by ID for sharing data through various print or web media and other Internet environments, comprising text, images, audio and video, and creating metamedia. Finally, ID has led to improvements in the fields of information presentation and representation, most evidently in the visible layer of a document, those surface aspects of graphical interfaces, such as the presentation of data and information, the definition of the typographic family, color patterns, and so on; However, it is important to point out that, supported by technological standards and codifications, ID has also had a huge impact on those aspects beyond the superficial, particularly in terms of interactive media projects, revealing objects and / or subjects in Internet information networks with greater clarity and enabling access to a large volume of information not only in traditional spaces but also in cyberspace, facilitating access, understanding and conversation between objects and subjects of information.
1.1. Disciplinarity and interdisciplinarity between KO and ID in Information Representation: a conceptual review

When studying interdisciplinarity in knowledge organization (KO), we find that the conceptualization and definition of ID is complex, given the many facets constituting it. Given that the process of designing and creating digital documents is inherent to knowledge organization in the contemporary world, in this context, we thus understand information design must necessarily be inserted in the new techno-social parameters – digital-virtual characteristics, supported by transversal competences – in the formation of a new generation of knowledge organization specialists, also known as curators.

On the other hand, the problems involved in the disciplinarity of KO take us to the roots of documentation, where we highlight, first, the classic definition of the document by Paul Otlet, as a generic term. In the scope of digitality, the document can also be understood as sets of documents of greater or lesser stability - such as those produced in ephemeral supports such as WhatsApp - but which always require, among other factors, that we take into account the social context of production and sharing. Otlet's ambitions for information socialization and access, and for all related actions are evident in his efforts to create a universal library with bibliographic records of world-indexed documents, using the format of a reference library, rather than of a collection (Oliveira 2005).

We understand that Otlet actually viewed his idea of a universal library as a platform (such as the Web 2.0), which also encouraged us to create and develop the project that generated this paper. Accordingly, in the genesis of our project, we responded to what Otlet defended and promoted as the invention or product of vernacular design with application to the transformations and revolutions of the digital age.

Information and communication technologies (ICT) have contributed to the development of graphic formats, which in terms of electronic support range from an abstract and symbolic interface (the written page) to a realistic iconic interface (the screen). With this in mind, we highlight the fact that Web 2.0 – offering graphic representations par excellence, as well as a platform of multimodal information – has, since the beginning of the 21st century, been an open, dynamic and continuously expanding complex system in which new applications and digital environments have regularly and exponentially appeared, just as Otlet envisaged for the universal library.

Over the last two centuries, information processing, organization and representation of documents’ information content has led to the development of standards, norms and protocols, allowing information to be represented through grammars and semantic models with computational connections, which has thus extended the reach and application of documentary languages. The adoption of thesauri, as a homogenizing representation of language terminologies, has become effective in controlling vocabularies and in solving the ambiguities of natural language. The functions of
thesauri in digital media can be expanded, to the classification of documents, the characterization of themes and categorization of concepts; the production and the translation of texts, mainly in the selection of vocabulary; communication and learning; the generation of the structure and conceptual basis for projects; the production of a conceptual context; decision-making in the classification of subjects; the summarization of texts; in the identification and association of key concepts; in information retrieval systems; and, finally, in navigation and labeling systems.

It is important to point out that the boundary between Web 1.0 and 2.0 was in the understanding of the Web as a platform (or as a reference library) in which ID has played a central role since the design of the new Web 2.0, to which various environments connect, transforming the platform into a dynamic, fluid and open system with multiple inputs. Interaction time, multimodal hybrid message structure, associative hypertextuality, interactivity and circularity, have all contributed to recreating information in Web 2.0, according to individual and unique perceptions and decision processes. Thus, the dynamic and fluid configurations of Web 2.0 have brought about a radical change, not only in message creation and new forms of knowledge, but also in the environments of health, custody and sharing of production and knowledge. New documentary languages have emerged in Web 2.0, such as folksonomies, which provide tagging or labeling of information in natural language through tags / hashtags that act as metadata (contextualized in communities of interest).

In Web 2.0, folksonomies are extracted from the participations of subjects who seek knowledge for information retrieval. Folksonomies can also be used in the development of controlled vocabularies with community-based (or bottom-up) indexing, constituting an efficient Web2.0 content classification system generated by netizens. Ontologies, on the other hand, are ways of representing and retrieving information in the digital environment. They are formal specifications that describe conceptual frameworks of specific domains, and result in a catalog of a domain of interest, created by specialists.

In this context, we find ourselves as specialists (or not) in documentary society (Buckland 2013, p. 9), in which the meaning of documentary processes is added to literacy and to orality, and any object that presents information can be an object of documentation. Bräscher and Café synthesized the characteristics presented by Fogl (1979) regarding information and knowledge concepts (Bräscher and Café 2010, p. 90), whereby information is a material form of the existence of knowledge; a definitive item of knowledge, expressed through natural language or other systems of signs perceived by the organs and senses; that exists and exerts its social function through a physical support; and which also exists objectively, outside the individual’s consciousness and independent of it, from the moment of its origin. Knowledge, on the other hand, according to the authors' synthesis, is the result of cognition (process of reflection of laws, properties of objects and phenomena of objective reality in human
consciousness); the ideal content of human consciousness.

For Capurro and Hjorland, knowledge and its communication are phenomena of all human society, but it was the emergence of ICT and its impacts that have characterized the information society (Capurro and Hjorland 2007), while information has become the main input of all sectors of society. For authors that advocate the paradigm of complexity (Morin 1986; Prigogine and Stengers 1984; Ferraz and Beltran 1999), on the other hand, technology is only one of the components of information systems, which are considered complex since they are composed of subsystems that interact and are interdependent (Jorente 2016).

1.2. Checking the occurrences of e-health Web2.0 platforms

The result of this exploratory phase of research methodology reveals the total absence of official digital environments directed towards non-specialist access to information on microcephaly in the context of Brazil.

2. Pursuing the challenge of immersive netnographic participation in the Whatsapp environment

In this context, design thinking, which combines with methodology from ID, is constituted of laboratory actions to solve or even redefine the problems initially traced. In a convergent way, for laboratory legitimation, we conducted a netnographic study, through authorized immersion in a closed WhatsApp group of mothers of children with microcephaly, named Mães de Anjos. Design thinking (DT) is a non-linear, human-centered methodology, which provides a solution-based approach to complex and unstructured problems that cannot be solved by conventional procedures based on linear thinking. Such an approach integrates observation (structured through the techniques of ethnographic / netnographic observation), collaboration and prototyping. Through the DT approach, an initial understanding of the problem is acknowledged, but aside from observation and experiments, other fronts and hypotheses can emerge from the peer-to-peer collaboration among the laboratory participants. When we performed the immersion, we verified, through the observation of natural language and its categorization, that needs are of different orders. We extracted folksonomy for the creation of controlled vocabularies. The 256 members of the WhatsApp group are mothers and grandmothers of children infected with Zika virus during pregnancy and who developed microcephaly and other neurological and musculoskeletal disorders.

Information, as a social phenomenon, depends on its communication, a process during which terminology can be extracted from records on informational platforms - natural language records of the netizens involved in the context that determined the

---

1 Considering that the research is documentary or representational, we opted for netnography for collecting the descriptors, and facet analysis for its categorization and vocabulary creation, to be applied in the digital platform in order to facilitate information search and retrieval.
information space, social construction and social living. From these records, data was collected, highlighting the most used terms related to the topic of microcephaly from the Zika virus. After extraction, a faceted analysis was conducted in order to treat and categorize the selected and collected terms and to identify key words (descriptors) used by the children’s families and related to the context of microcephaly, based on theoretical and practical principles consolidated by the theory of faceted classification.

The colon classification theory, developed by Ranganathan in the 1930s, presented itself as an alternative in the flexibilization of the rigid ordering of hierarchies, contributing to improve knowledge sharing among the participants. The categorization provided a more comprehensive set of concepts and their organization in classes and sub-classes, essential for information analysis, synthesis and representation, offering the basis for information search and retrieval in an informational environment. Four categories of subjects shared by the community of interest were extracted: a) medicine; b) treatments; c) clinical status; d) benefits of government / public institutions.

Within these categories, a large number of subjects of interest demand the urgent action of professionals involved in multidisciplinary teams to create environments for sharing knowledge about the consequences of the Zika virus for women who, as a consequence, have given birth to children with the resulting syndromes of microcephaly.

Partial conclusions
The observation systematized by DT in netnography showed that the descriptive information among peers about clinical status was performed in simple, non-technical language and natural language (NL). It was possible to identify keywords, which were collected, treated and used as descriptors for indexing and for the formulation of queries for effective information retrieval, since most people, when searching for information, use natural language, from which the content creation process must be inferred (with a vocabulary closer to their informational, cognitive and contextual realities and needs). The results sought in this stage were related to the identification of keywords used by the children's relatives, about their health conditions and informational needs.

Since this is a work in progress, the actions in the implementation phase are: a) the creation of controlled vocabularies from verified folksonomy; b) the study of visual elements for the translation of complex knowledge (intersemioses); c) the final recommendations for the creation of a digital environment model, for organizing knowledge about Zika virus and microcephaly.

The next phase, interdisciplinarity with computer science will also propel development of applications, software, tools and digital environments to enable tasks to be carried out taking into account mobility and portability.

This research, justified by the need for academic-scientific action for knowledge
organization in e-Health digital environments, is based on the assumption of the need for information sharing and exchange to empower netizens, whether specialized or not. The strategic use of ICT is a premise that allows the participation, training, education and work of non-specialized individuals to assert their basic human rights to access and share knowledge production and information related to their health and living conditions. We advocate that models of digital environments in the area should also provide conditions for the affected people to share ideas, experiences and questions. However, to achieve the effective sharing that we propose, the system must be planned top-down, provided with an ID that stimulates the emergence of information and spaces for knowledge exchange among the various social actors, in a bottom-up or horizontal fashion.

References


Intersectionality and the social construction of Knowledge Organization

Abstract
The aim of this paper is to identify, systematize and analyse the political and conceptual needs that guide the composition and update of knowledge organization devices in contemporary times, along with their ties to socio-historical conditions and ethical modes of existence. The analysis took as a reference the intersectional processes of representation of knowledge in the themes of gender, race, sexuality and feminist studies in distributed and networking information systems.

Introduction

The processes of information and knowledge production, circulation and organization have undergone several transformations in the last few years. We can readily see a rise in intense questioning about the limits for knowledge organization systems which, from the perspective of pervasive criticism, tend to reify knowledge for a shared verisimilitude within hegemonic groups. Such procedures would lead, ultimately, to the production of a partial mirroring of reality, rendering the presence of some themes, representations and subjectivities rarefied.

Accordingly, approaches calling for a transversal and intersectional turn based on devices structured in the context of knowledge organization are growing in numbers.

Considering the context of socio-technical changes and shifting power relations, the concerns and public perception that knowledge organization devices can be used as tools for domination systems has become more widespread (Berman 1971; Olson 1998; Drumm 2000; Mai 2010; Higgins 2016; Adler 2016).

Consequently, we are at a social-historical moment in which the theme of discursive reparations in communication and knowledge organization domains finds room for reflection.

Foucault (1995) emphasizes that contemporary struggles are transversal, targeting the effects of power; these struggles are no longer limited by geographical boundaries.

Through an analysis of power relations, Foucault (1995) identifies a triple typology marked by support, imbrication and functioning as an instrument. They are as follows: communication relations, power relations and objective capacities. The author highlights that these relations are neither uniform, nor constant. However, “there are also ‘blockages’ in which the adjustment of capacities, the bundles of communication and power relations constitute regulated systems and accords” (Foucault 1995, 241). Disciplines, then, have their role under the spotlight because they “show, according to
artificially clear and decanted schemes, the way objective finality systems, communications and power systems may articulate over each other”.

Foucault understands that power is “a way for some to structure the field of possible actions by others”. For that reason, it was considered pertinent for this work to reflect on the unfolding of such a perspective in the composition and update of knowledge organization devices. It was also perceived that it is relevant to identify remains kept by these devices with socio-historical conditions and ethical modes of existence presumed in the discourses that promulgate them.

To achieve that, we adopted as reference the methodological principles used by Foucault to approach issues of power, particularly in regard to his archaeology of knowledge, which focuses on discursive formation entering the genealogy, in which the philosopher may integrate non-discursive relationships and acquire knowledge through strategies and tactics of power. In that aspect, we believe that it is possible to understand the workings of interlacement of discursive and non-discursive practices in relation to power and knowledge in informational contexts and specific devices.

Foucault’s genealogy is an attempt to de-subject historical knowledge and aims at activating local, discontinuous, unqualified, unlegitimized knowledge against the unitary theoretical instance, intending to debug and set them in a given hierarchy in the name of a single true knowledge.

From that perspective, Foucault’s genealogy encompasses:

- A historical ontology (us – ourselves in our relationship with truth – us as knowledge subjects);
- Our relationships with the field of power (subjects acting over others);
- Our relations with morals (allowing us to construe ourselves as ethical subjects).

In the development of this study, we considered Foucault’s guidance regarding power relations, seeking to describe, in the field of information organization, the distinction between objective capabilities and communication relations. The objective capacities involve power exerted over things, that is, the capacity to modify, use, consume and destroy such elements. Information/communication relations involve interactions linked to transmission processes in symbolic systems.

Intersectionality is a concept proposed by Crenshaw in 1989 that refers to the multiple interactions that constitute the human being. These interactions are invariably crossed and informed by different systems and power structures. From a structural perspective, intersectionality reveals schemata and interfaces that prevent egalitarian access to power structures and, from the political point of view, emphasizes the perpetuity of the marginalization of certain social strata.
According to Crenshaw (1993, 1.242),

Race, gender, and other identity categories are most often treated in mainstream liberal discourse as vestiges of bias or domination – that is, as intrinsically negative frameworks in which social power works to exclude or marginalize those who are different. According to this understanding, our liberatory objective should be to empty such categories of any social significance.

Lutz (2015) claims that concept of intersectionality can be considered a heuristic device or method.

The perspective of adopting intersectionality as a method to study social inequalities can, according to Lutz (2015, 43), avoid certain traps. In this context it is necessary to question positions of privilege. In that sense, Floya Anthias (1998 apud Lutz 2015, 40) proposed a “multi-level analysis that works on four levels: the level of discrimination (experience); the actor level (intersubjective praxis); the institutional level (institutional regimes); and the level of representation (symbolic and discursive)”.

Lutz (2015, 40) proposes that, before information which seems to be “about gender”, it is necessary to ask about "other questions" that regularly involve the reflexion about the subject "exposing multiple positions and power inequalities", for example, class interests, patriarchy and sexism.

In the sphere of knowledge organization processes, guarantees work as analytical operators guiding the methods for knowledge organization systems (KOS). Historically, these operators have been under suspicion due to their naturalization
function and its risk for perpetuating prejudices, absences and regimes of oppression they may engineer.

In that sense, semantic guarantees, as Campbell (2008) *apud* Roger (2016, 110) claims, reveal a “permanent commitment from systems and information contexts for analysing, justifying and reviewing semantic elements that structure the represented knowledge”.

According to Fox (2016, 582),

Epistemology, the study of knowledge, examines who can “know”, who “creates” knowledge, what authority and interests they possess and what it means to how a concept is understood. KO research took an epistemic turn in the mid-1990s, as researchers began to recognize the social influences on classification and began to examine how knowledge was legitimised for classification purposes. (...) A relative of epistemology is ontology or the study of what “is” or “exists”, and specification for determining something that exists. (...) Here I refer to it in somewhat of hybrid sense in that if epistemology reveals how people decide how knowledge is legitimised at specific points in time, ontology provides embodiment of that knowledge, whether through literature or other means.

In this study, our effort is to try to understand the possibilities of the intersectional method for consolidating and updating KOS considering the interface between epistemology/ontology and social context.

**Objectives**

The study aimed at analysing the relations between intersectionality, power relations and language as theoretical scales of explication and knowledge organization device updates. It sought to rethink the analytical assumption represented by canonical warranties (literary, of use, and structural) historically used as inhibitors for language subjectivation in situations of knowledge representation. In order to achieve these aims, a KOS was built, guided by the intersectional perspective in the themes of race, gender, sexuality and feminist studies and based on specialized literature and the language used in spaces for sharing information, focusing on social activism on those themes.

The KOS created was compared with the speeches from the SlutWalk Movement in Brasília (Brazil) and the Women's March in Washington (EUA) with the objective of identifying limits and possibilities for intersectional representation in the proposed themes.

**Methodology**

In this study, the potential for updating KOS in interface with the information and knowledge production flow was analysed along with its ability to capture and represent the multiple discursive disputes that are established in the context of knowledge production. For that, the speeches about the transnational movement, SlutWalk, and the Women's March on Washington of 2017 were analysed.

We sought to understand the main concepts adopted by these movements and their
main distinctions. Wes also sought to understand the intersectionality within the discursive production of these movements, having, as reference, the multilevel analysis proposed by Anthias (1998).

The SlutWalk Movement, established in Toronto (Canada) in 2011, was motivated by protests against the direct connection between the dress code and violence against women. In the context of the emergence of SlutWalk Movement, there were a number of sexual abuses against women at the University of Toronto. At the time, the police officer Michael Sanguinetti said that women could avoid such violence if they did not dress like whores, which gave rise to numerous protests internationally.

At the first march, held on April 3, 2011, about 3 thousand people took to the streets. According to Carr (2013, 21-28),

SlutWalks represent the spontaneous outrage of women, the LGBT communities, and pro-feminist men around the world against patriarchal rhetoric (...). The SlutWalk movement presents an opportunity for scholars to apply feminist theories to a new form of transnational feminist activism located at the margins of mainstream society, drawing upon transnational feminist and social-movement theories.

The Women's March on Washington, held on January 21, 2017, the day after the inauguration of Donald Trump, thematized an intersectional social agenda that articulated the defence of human rights, the fight against racial inequalities, the strengthening of labor rights and the reduction of environmental impacts.

The manifesto of the Women's March on Washington claims that,

The Women’s March on Washington includes leaders of organizations and communities that have been building the foundation for social progress for generations. We welcome vibrant collaboration and honor the legacy of the movements before us – the suffragists and abolitionists, the Civil Rights Movement, the feminist movement, the American Indian Movement, Occupy Wall Street, Marriage Equality, Black Lives Matter, and more – by employing a decentralized, leader – full structure and focusing on an ambitious, fundamental and comprehensive agenda.

Table 1: Comparison between the speeches from The SlutWalk movement and Women's March on Washington – according to the Floya Anthias Model

<table>
<thead>
<tr>
<th>LEVELS</th>
<th>SLUTWALK MOVEMENT (1)</th>
<th>WOMEN'S MARCH ON WASHINGTON (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrimination (experience)</td>
<td>Protest against sexual violence; in favor of sex positivity and sexual empowerment,  -We have been called sluts because we wore short clothes, we were called sluts because we have already made love before the wedding, we were called sluts for simply saying &quot;no&quot; to a man, we were already called sluts because we raised the tone of voice in a discussion, we were called sluts because we did not follow what the society or our family expected of us.</td>
<td>Black women, Native women, poor women, immigrant women, Muslim women, queer and trans women</td>
</tr>
<tr>
<td>Intersubjective praxis</td>
<td>Gender Justice</td>
<td>Racial Justice</td>
</tr>
<tr>
<td>------------------------</td>
<td>---------------</td>
<td>---------------</td>
</tr>
<tr>
<td>feeling unsafe; slut- shaming; - We march for the health of black women, because we have less access to health services;</td>
<td>- Safe and healthy environments free from structural impediments - we must protect and restore all the constitutionally mandated rights to all our citizens” - We believe Civil Rights are our birth right. - Immigration reform must establish a roadmap to citizenship - We believe it is time for an all-inclusive Equal Rights Amendment to the U.S. Constitution. We believe in immigrant and refugee rights regardless of status or country of origin.</td>
<td></td>
</tr>
</tbody>
</table>

| Institutional level | - we marched because there were about 684 police inquiries into rape crimes - an average of two women raped each day, -We march because many of us depend on the precarious public transportation system of the Federal District, which forces us to walk long distances without any security or lighting to protect the various women who are sexually abused along these routes. - We will continue marching because we live in a patriarchal culture that activates diverse devices to repress the sexuality of the woman, dividing us into "saints” and "whores. - We will continue to march because women are still a minority in positions of power and receive on average 70% of men's wages. We will continue to march because there are jobs performed by a female majority that are not recognized or endowed with economic value because domestic workers are invisibilized, exploited, discriminated against and have not secured some of the most basic fundamental rights of work. - We will continue marching because prostitutes are part of the functioning of a macho and hypocritical society that, while using their bodies, insists on denying their citizenship. |

| Representation (symbolic and discursive) | - Recognizing that women have intersecting identities” - Our liberation is bound in each other’s -We are empowered by the legions of revolutionary leaders -We believe that Women’s Rights are Human Rights and Human Rights are Women’s Rights” We believe in accountability and justice for police brutality and ending racial profiling and targeting of communities of color. -We believe that our environment and our climate must be protected, and that our land and natural resources cannot be exploited for |
|------------------------|------------------------------------------------|------------------------------------------------|------------------------------------------------|
| - We are sluts: feminist women, who are struggling against discriminations of race, sexuality, gender, religion and class. We are entitled to our lives, and to live them as we wish, and free from any form of violence. - We believe that the end of violence against women is directly linked to the transformation of conservative and hegemonic values in our society, as well as the overcoming of patriarchy, all fundamentalisms, lesbophobia, biphobia, transphobia, homophobia, machismo (chauvinism), racism and the capitalism. - We argue that we all have the right to choose about our bodies. | - We are entitled to our lives, and to live them as we wish, and free from any form of violence. - We believe that the end of violence against women is directly linked to the transformation of conservative and hegemonic values in our society, as well as the overcoming of patriarchy, all fundamentalisms, lesbophobia, biphobia, transphobia, homophobia, machismo (chauvinism), racism and the capitalism. - We argue that we all have the right to choose about our bodies. |
Main results

Considering that power and oppression regimes present a multiform perspective that is often naturalized as language, the work resulted in the understanding of possibilities for articulation and increasing compatibility of cultural warranties as an umbrella concept (Guedes 2016, 90), incorporating other warranties. From that point of view, a dialogical and intersectional KOS was consolidated, considering in its formulation the fundamental role of sensibility concepts (Guedes 2016, 114-115), such as ideology, culture, intersectionality, power and ethics. From the perspective of modelling and language agency, an attempt was made at demonstrating, through “COEXISTENCE | Thesaurus of intersectionality | race| gender| sexuality | feminist studies”. The Thesaurus is based on diverse layers that compose identity politics in the social sphere and the possibilities for critical modelling of the language derived from these practices.

Conclusions

Understanding the socio-cultural, arbitrary and plural characteristics of language makes explicit an increasing demand for an interdisciplinary dialogue and a return to theoretical and experimental issues previously effected in the field of knowledge organization.

We believe that the use of the multilevel analysis method, proposed by Anthias (1998), allowed us to create this thesaurus and to identify the intersectional themes related to the analysed social problematic. In this way, it was possible to reveal, in discursive terms, the contradictions, the regimes of oppression and the positions of privilege in the context of the social struggles fought by contemporary feminist movements.

We conclude that the elaboration and consistent application of KOS results from discursive disputes in the socio-cultural context presented as semantic relations mediated by socio-technical devices.

In this study, the translation of a multiform dynamics of contemporary identity processes in KOS required the acknowledgement of the historical effort done by researchers from the field of knowledge organization, followed by an admission of the need for new forms of preventing crystallization in infinite discursive disputes, existing within knowledge representation systems. In these terms, the concepts of power and structural and political intersectionality provided an understanding of the multidimensional character and the reciprocal influence that are present in social
relations mediated by the combined presences of racism, sexism and patriarchalism. This allowed for a reduction in the effects of an essentialist view of the concept of woman and social hierarchies that result from it, in the experience of creating a dialogic and intersectional KOS.

Acknowledgements

Acknowledgements are owed to CNPq and Fapemig, for the support provided during the different stages of development for this work.

References


Knowledge Organization in the digital age: the complexity of the global labor market

Abstract

In this paper, we focus on strategic aspects of the new profile of knowledge organization professionals. First, we discuss the dimensions of knowledge organization in a current globalized and competitive scenario as an inherent characteristic in any organizational action, a vital element in the production of social reality. Secondly, we highlight the skills of knowledge organization professionals in the field of new technologies and knowledge distribution. Thirdly, we approach the restructuring of the organizational environment considering the perspective of the global labor market complexity. Finally, we present our reflections on the skills and competences of the knowledge organization professional with respect to the global labor market’s expectations and demands in the digital age.

Introduction

The strategic value of science, technology and innovation has been accepted unquestionably. Improvements in these areas act like a fundamental driving force of industrialized countries and they are also responsible for enhancing their wealth. In recent years, industrialized countries like Singapore, China and South Korea, have revolutionized their economies; they are now modern and extremely dynamic. Even when following specific political tracks, it has been no different in those cases: the incentive and dissemination of a culture based on promotion of information, knowledge and technological innovation. Technological development represents an indispensable condition to effectively compete for best levels of position in crowded global markets.

In recent years, information literacy has constituted a new research field that corresponds to education, information science, knowledge organization, and cognitive science. First, the challenge is to learn the basic use of technological resources - digital literacy - and second to adapt and generate new knowledge - information literacy. In the information literacy context, the challenges are multiple: the difficulties of university libraries with their expensive collections to compete with digital collections; virtual libraries and search engines providing information instantaneously at no cost; the deconstruction of the individual author concept and the emergence of digital collectives.

The recent and rapid development of emerging countries, grounded in a culture that promotes information, knowledge and innovation, highlights the importance of education as an economic element to further wealth generation and income distribution. In the globalized scenario in which knowledge has a strategic role in the definition of management policies focused on organizations’ competitive advantage, information
science plays a strategic role in the development of a country's competitiveness.

The advent of rapid knowledge obsolescence also reflects the need for professionals with holistic vision, management skills, methodological approach, cultural and systemic vision, creativity, entrepreneurship, innovation, environmental responsibility, communication and leadership. Global competitiveness requires a new professional profile, which can face the challenge of balancing the skills of a strong academic and technical background and production process management, focusing on global competitiveness where information has strategic value in decision making. This situation requires educational and professional skills that meet the demands of the global labor market. In addition to an excellent humanistic, cultural and technical formal education, organizations need professionals with new competencies and abilities.

The network society’s new configuration constitutes the object of study and research for professionals from various fields, especially information workers, who deal with the challenges of information management and knowledge organization in their dimensions (Figure 1): scientific, technological, industrial, marketing, strategic and more recently, social.

In this new context, one of the most imperative challenges of education in science and information technology, where technology plays a central and decisive role in information organization and management, as well as the production of new knowledge, is to develop curricula that provide training aligned with the challenges of the global marketplace and, at the same time, considering the cultural and humanistic education of professionals that are working in the area, aiming for the training of leaders that can actively contribute to sustainable development and wealth generation (Observatório 2011).

In the process of training professionals with globalized skills, a relevant factor that
should be considered by universities in shaping their faculties is to consider not only academics with PhDs and Master’s, but also professionals who hold active and strategic positions in the work market, thus enriching the academic environment with new experiences, technologies and trends. Lessons, both theoretical and practical, should make use of modern computational resources to enable renewal of the research environment, development and production of the productive sector and research, leading the student and teacher to experience the competitive environment in which organizations are located (Brascom 2018).

To face international competitiveness, we need to reappraise the strategic value of information and knowledge as well as understand that there is still a lot of work to be done in this field. There is a connection between technological development and sustainable economic growth that is essential and involves many areas. Among them, one of great interest deals with the modernization and internationalization of the current academic model. It is not enough to ensure good student training, we need to develop new skills required by global labor markets.

In this context of increasingly dynamic change, knowledge becomes obsolete quickly. In the case of information science, which is cutting edge compared to many fields of scientific and technological knowledge, it is possible to say that half of what the student learns at university will be obsolete in five years. It is recommended to organize a holistic qualification, enhancing management, communication, leadership, methodological, cultural, multidisciplinary and systemic skills - all of them highlighted in the knowledge economy.

We need to understand and appreciate the complexity of the contemporary world to face the challenges of education. In addition to specific technical skills - indispensable in information science, most new or renewed professions will require the practice of many cultural abilities. To educate the information professional for the 21st century is to balance the binomial of expert - in its technical dimension - versus generalist; in short, by adopting a multidisciplinary approach.

**Education and professional skills in knowledge organization**

In a world where to achieve higher reaches indicates the measure of organization survival, to remain oriented to a more globalized market, despite all the economic and structural difficulties, is not an easy task. The focus must always be on the future. We must dedicate ourselves to build an innovative educational project, preparing the new information and knowledge professional to accept the challenge of exploring the path of innovation and sustainable growth. The best performing organizations are those able to identify opportunities, take advantage of the changes, turn ideas into reality and get the results that put them one step ahead of the competition, ahead of their time.

Keeping in mind the vital importance of the role of information and knowledge in
defining the competitive strategies of organizations, establishing methodologies for its organization, management, access and use is the bridge between science and society. Education in this field of knowledge must be based on three pillars: technical training, cultural and humanistic education, and environmental responsibility for sustainable development.

Of crucial importance, the lessons of information science courses, which are preparing these new professionals, should develop, in their pedagogical projects, the principles of sustainable development, focusing beyond the immediate future, with innovation and creativity seeking balanced solutions between the needs of society and the effects on the environment. Science, information and technology are strongly interconnected, and it is important to have a better comprehension of how to manage access and use of information in converting the new scientific knowledge into a technology at the service of modernity (Paletta, Maldonado 2014).

In a world without barriers to knowledge production, "mobility" has become a key concept for every professional and for all organizations that compete in an increasingly globalized market. Mobility must be understood not only in its physical aspect – because in a world dominated by information and communication technology, mobility is becoming increasingly "virtual" - but especially in the sense of flexibility, adaptability and interactivity (Canclini 2009).

Mobility is the set of attributes that enables one professional to seize new opportunities, either in foreign countries or at the place of origin. Mobility requires skills that go beyond traditional academic training, and the guarantee offered by international standards certification and accreditation of higher education diplomas.

This is an irreversible trend that results in new forms of organization of production on a global scale, examples of which are outsourcing or third part services within national borders; the offshore or international third-party services; and the formation of supply chains and information. Mobility imposed by the need to guarantee competitiveness in regional economic blocks, as well as local development, in response to the efforts of global competitiveness.

To achieve this mobility, the information and knowledge worker needs to combine traditional technical and scientific knowledge – basic elements of information science – with management skills that qualify him to assume responsibilities in the new organizational environment and the need for the spread of organized and productive information.

The development of information technology has followed the course of the industrialization process. In the first place, the required competence was eminently technical. In a second stage, as the industry diversified and became more sophisticated, scientific qualifications were required. In a third step, managerial skills were then required. The direction followed in this process was one of increasing specialization.
Progress was then made to a fourth stage, which was reached by the reverse direction going from expertise to holistic training (Paletta 2012).

Holistic training as a requirement of mobility, is related to mental flexibility and, therefore, with a kind of innovation. The relationship between holistic knowledge, globalized markets, the knowledge economy and sustainable development is intrinsic.

For this global professional, to have holistic competency means adding new knowledge and ability to basic technical skills, such that this professional should live together with communities and cultures that are living and solving daily issues and problems considering a specific and characteristic point of view. The professional should have communication skills and know how to work in multidisciplinary teams and be aware of social, ecological and ethical implications involved in the management, access and use of information and knowledge production.

Another important point is to speak more than one language and be able to work anywhere in the world. A compilation of recent studies sums up the type of expertise and skills required today in a global professional:

- performance in multidisciplinary teams;
- identification, formulation and solution of knowledge management problems;
- sense of ethics and professional responsibility;
- recognition of the need for continuous training;
- use of modern techniques and tools to have good search practices, access, ownership and use of information;
- informed socio-environmental responsibility;
- understanding of ICT solutions’ impact in a global and social context.

The formation of such skills requires that the technical disciplines provided in the curriculum guidelines of university classes should be complemented with interdisciplinary content, and to the theory to be attached to problem solving. In this case, cooperation between universities and organizations is fundamental. Understanding the historical context in which they develop information science in different countries contributes to breaking down cultural barriers.

The complexity of the global labor market

Knowledge is the factor that currently adds more value to people, services and products, in any area or professional field. For this reason, qualification should be the training goal of information science professionals. Qualification is understood as a logical result of a process that depends on what is loaded in the brain and not the mere enhancement of information without connections or contexts, memorization without critical consideration. To the extent that the world is sophisticated and diversified, the competence required of the information professional that can be considered highly technical - remembering it is impossible to dispose of it or move forward without its
consolidation - widens and encompasses other important responsibilities in the new organizational environment.

Thus, it is necessary to think in a holistic qualification aligned to the demands of the knowledge economy. The knowledge economy can be highlighted in reference to the cognitive and pedagogical models able to dialogue with the ever-changing world. Nowadays, intelligence has another meaning. It means to perceive things that have not yet been seen. It indicates an extensive qualification. It is a qualification that teaches students to think, catalyzing several multidisciplinary and transdisciplinary projects, valuing curiosity, the ability to explore, to solve new problems, be constantly updated, and develop a leadership profile, requiring the autonomous pursuit of knowledge on the complexity of the world.

In two decades of the new millennium, the development of organizations relies on technology from both a social and economic point of view. Technology is a decisive factor in the digital divide and wealthy generation issues, based on new products, services and processes, all of them heavily dependent on innovation. The ability of an organization to act in a global market depends on a competitive positioning, as the creator and developer of technology-based products and solutions. Therefore, in the coming years emphasis should be given to technology-based entrepreneurship, that is, the ability to offer new products and services based on innovative technologies to the market.

This challenging role is up to all sectors of society and to information science professionals, who should be able to organize and manage information with a focus on producing new knowledge, using new techniques, creativity and audacity to offer global society solutions with differentiated content that can improve not only the productivity of organizations, but also the quality of life. This entrepreneurial mindset, not purely commercial, is a key feature in the development of professional skills with ability to create wealth through technology and knowledge.

The influence of science, information and technology is evident in organizations’ abilities to compete aimed at achieving sustainable development of society within a well-defined global positioning. It is up to the university, in technological cooperation with the production and services sector, to contribute to the proper training of human resources, offering not only technical training but also humanistic and global expertise, so that new information professionals can become wealth-producing vectors of income distribution and sustainable economic development.

To consider the education mission, it is necessary to comprehend and appreciate the complexity in contemporaneity. The future intelligence and the future of intelligence depend on the acceptance of complexity and how we deal with it. It is important to consider dialogue in any educational project. It is possible to say that we are bound to complexity and it will respond by guaranteeing future advances in medicine, genetics,
robots, artificial intelligence, energy, environment, pharmacology, nanotechnology, computing, agriculture and of course all areas of information science. Jacques Attali (2001), in his dictionary of the 21st century, says, "According to Moore's Law, the complexity of electronic chip increases 50% in eighteen months. According to Metcalfe law, the value of a network increases with the square of the number of those who use it. According to Kao's law, the creativity of a group increases exponentially with the diversity and divergence of which it is composed." Therefore, complexity will increase in consequence of the human adventure during this century (Paletta, Maldonado 2014).

Complexity is the great gift that life brought to our planet. Of course, a correspondence can be established between the complexity of the human nervous system and the complexity of human knowledge. Like the world around us, neurons are extremely complex: they are numerous, multiple and engage in several different ways, creating a web of diversity. Maturana and Varela (2000), in The Tree of Knowledge, understand that a nervous system is just as rich and vast as the man is - "tens of billions of cells in a combination of truly astronomical possible interactions" - provides new dimensions of complexity, responsible for language and self-consciousness; in short, the raw material to be treated by education. Either we learn to deal with complexity or paradoxically, the result will be chaos.

Complexity renews the concept of education, which is intrinsically problematic because it is related to an activity that transforms and creates alternatives to the usual. If the world is infinite and, more interesting, "infinite in all directions", education needs to be even more open to possibilities, expanding its versatility (Freeman 2000).

The philosopher Pierre Lévy (1998), professor of the chair of Research in Collective Intelligence at the University of Ottawa, reaffirms the importance of education to the world's complexity. He observes: "In every place to where we direct our gaze with sufficient accuracy and perseverance, the world in which we live shows its complexity". The possibilities opened by a qualification that takes complexity into account are ample and appropriate to the future scenario of professional work.

The dynamics of most professional activity in the next decade are not clearly established. Professions will continue diversifying and new professions that do not yet exist will be in great demand. In addition to specific technical skills – indispensable, in the case of information science – most new or renewed professions require the practice of many cultural capabilities. To educate the information professional for the 21st century is to balance the binomial of expert (technical dimension) versus generalist (qualification for complexity), in a smart way because the future possibilities in the labor market are undoubtedly smart (Lévy 1997).

Conclusion

Knowledge organization links three processes of strategic information use – the
creation of meanings, knowledge construction and decision making – a continuous cycle of learning and adaptation that can be named a knowledge cycle. In recent years, the value of science, technology and innovation has established itself as strategically unquestionable. Advances in these fields act as the driving force of industrialized countries, in response expanding their wealth.

In this context, in France we can highlight the Professional Association of Librarians and Professional Information, established in 1963 as “Association des Bibliothécaires et Documentalistes Spécialisés” and which in 1993, in recognition of the expansion of the profession scope, was renamed to “Association des Professionnels de l'Information et de la Documentation”. The Association aims to promote the profession in its various forms and develop information competencies of professional interest. According to the L’Observatoire de la Fonction Information (2018), the activities related to information organization and management are developed in a complex and challenging ecosystem. Different actors coexist in perimeters of action throughout the chain of information services. In a network society, it is imperative that information professionals work in cooperation with several fields of knowledge and specialized activities considering technical and methodological expertise.

Scientific knowledge and access to technological innovations are very unevenly distributed considering the comparison between countries, regions, social class, age groups, educational levels, among others. Thus, the issue of cultural diversity and research about it, should be part of the theoretical consideration, of empirical research and policy planning in the information area of literacy development. In addition, the typical information professional does not want only to extract specific and definitive information as soon as possible or, alternatively, the professional is willing to invest effort to seek and exploit information and knowledge. The truth is that people continually oscillate between extract and exploit, and use of information is a confused process, disordered, subject to fluctuations of human nature, like any other activity. Among the most important elements that influence the use of information are the individual's attitudes toward information and its search. To the extent that the world is becoming more complex and diversified, the skills required of information professionals, in a first stage, are a highly technical flare which will come to encompass other responsibilities. It is necessary to consider a holistic qualification, enhancing managerial, methodological, cultural, multidisciplinary and systemic skills. However, the role of technical training should not be minimized.

The professional must update his/her knowledge and skills constantly since the technology becomes obsolete so quickly. This process involves not only ensuring the proper technical training for students, but also finding ways to develop new skills such as management, communication and leadership because these qualities are increasingly demanded by the labor market.
Acknowledgment: FAPESP Research Project – Processo 2016/07358-6

References


Samuel Ayobami Akinruli, Luana Carla Martins Campos Akinruli

Memories in dispute, and reconfigurations of cultural heritage: for an Ethnography of museums

Abstract
This paper proposes possibilities for ethnographic practices in the context of cultural heritage studies, focusing on the museological collection, taking as a case study an ethnography of the practices involved in the musealization of objects in the museum called “Museu Casa de Lembranças e Memórias Chico Xavier”, located in the city of Uberaba – Minas Gerais, Brazil. In order to do so, we seek to promote an interdisciplinary discussion between the fields of anthropology, ethnography, history, museology and museography, which is based on the dialogue between the materiality and immateriality of collections and practices of organization and museum inventory, in order to identify the forms of interaction, reflection, communication and social representations of cultural heritage. The balance of the institution’s founding process, of the composition of the collection and of its inventory process, interacts with different levels of temporality and allows dialogue between the construction process of certain memories that are in open dispute with various agents and agencies of society. The purpose of this approach is to discuss the association of cultural inventory and diagnosis with the implementation of security, protection, safeguarding and broad access to information and the collection of both material and immaterial assets, based on the convergence of some specific methodologies in the field of human sciences, especially associated with cultural heritage, that integrate in a dialogic way the large volume of information related to the cultural heritage of intangible and tangible nature. It was considered that in order to obtain effective knowledge, identification, preservation and management of the collection, and taking its huge size into account, the volume and scope of cultural heritage as a conjunction of objects, the ethnographic basis should thus be the point of connection between the applicable methodologies used by the regulatory organs at the municipal, state and federal levels. Ethnography makes us question not only the themes, approaches and description methods, but the way of constructing the narrative itself, the agencies that are not only human, the supposed separation of theory and practice, as well as the field and ethnographic writing. To this end, questions are raised about the sociological unit in its research reality, in a way that the displacement of the category group / collective, private / public; material / immaterial occurs. Experimental and ethnographic writing, when problematized, call into question the very mechanism of research and the model of scientific production – including that of anthropologists – based on a discourse of authority. It is not just about style, semantic choices, or textual formats, but ethnography promotes an invitation to think about the researcher's performance.

1. Memories, cultural heritage and social representations: contributions to Ethnography of museums

1.1. Heritage, memories and social representations
Cultural heritage has attracted the interest of a significant number of researchers, principally among them museologists, anthropologists, architects, archaeologists, geographers and historians, who, through different approaches and themes, contribute to the formation of a broad set of varied studies, many of which focus on the research and analysis of cultural phenomena.

In Brazil, the discussion of heritage gained momentum with the first preservationist
movements led by members of the Modernist Movement who also acted in the newly created body for preservation of national historical and artistic patrimony: The National Historic and Artistic Patrimony Service (Serviço do Patrimônio Histórico e Artístico Nacional – SPHAN) in 1937. In the search for Brazilian traditions, a national identity was constructed, represented by the majestic baroque churches and works of masters such as Aleijadinho. Moreover, the manifestations of popular and indigenous culture, such as music, tales, legends, medicine, cuisine, among others, were mentioned in the preliminary draft for the Protection of the National Artistic Patrimony produced by Mário de Andrade in the 1930s.

It is also important to underline the great social transformations generated by industrialization, especially between the mid-19th and 20th centuries. The term heritage assumed a social and even symbolic relevance in this context, being affiliated with preservationist movements as it attests to the process of modification, loss and/or destruction of cultural collections. This assertion is still relevant today in the context of the great impacts generated by the economic expansionist model.

The differences and heterogeneities of culture as a strategy for the construction of nationality have been eliminated. There is in this process a latent dichotomy between two conceptions of seeing and living in the world that are inherent to the context of modernity: the past is often regarded as something that needs to be overcome in search of a new civilization, while at other times, it needs to be remembered, given the accelerating process towards amnesia.

One of the issues raised in this project is that, in recent decades, the use and social functions of heritage have expanded the concept as well as broadening the need to interpret its dynamics in contemporary society. With wide-ranging interest in the expected benefits, and in a context of accelerated speculation regarding cultural assets, we consider it important to highlight the social relevance of cultural heritage. Social values inform analysis of how equity is interpreted over time, as well as how to organize classification and promote safeguard policies.

Postcolonial studies, by their nature, promote the situation in which communities manifest themselves as decision-makers in defining what is considered as heritage. New collectivities become visible in the social fabric, seeking to assert their identities whereas longstanding privileged identities which were up to then hegemonic become questionable. The nation is seen not only through the prism of homogeneity, but through the various differences and multiplicities present across its territory. These dynamics are the basis for appreciating local cultural references, of a material and immaterial nature, in order to seek knowledge of a history open to diversity and paths open to the constitution of cultural citizenship.¹

¹ What is called cultural citizenship depends to a great extent on interpreting the fundamental cultural and
From this historical perspective, understanding the notion of cultural heritage depends on several interlocking factors. However, such premises are not enough to fully apprehend the problems raised by this doctoral research, since the assumption is that cultural heritage has very strategic uses in contemporary times, with very specific political and economic functions. What are the attributes and reference values for cultural assets to be considered heritage? What cultural references should be included or excluded from inventories? Of whose interest is it to preserve or to forget?  

The memory that celebrates cultural heritage is closely linked to certain interests and desires to build a certain representation and a collective memory of social cohesion. All that is real, understood as knowledge and interpretation that men attribute to reality, is a representation of what actually existed, and history, as well as memory, also construct representations of that past, so that the “representation of an object then corresponds to a set of information, opinions, and beliefs concerning this object”.

Thus, the identity of a group – its peculiarities, similarities and differences from other groups – is constructed through a set of representations, images and symbologies embraced with meanings, by memory. Memory, therefore, works actively in the reconstruction of the past through its previous experiences, and interferes in the means of interpreting realities.

The lack of understanding of specific strategies of memory repertoire is also one of the results of modernity that tends to detrimentally undo the old for the new, generating waves of nostalgia represented by a desire to safeguard the lost time. The sensation is one of loss, of an emptiness that can only be filled by the mythical admiration of the old thing, by the irresistible attraction to historic centers and past constructions, and by the increased tourism provoked by museums. However, “what is more important, in this case, is the reflection of the implantation of the museum on the life and growth of
the city. The content of its collection is intrinsically linked to the memory of the community”.

Museums reveal the tensions and contradictions between the need to forget and the desire to remember, but museological procedures are marked by very specific selection criteria determining what is cut out. Preservation is the reaction to the extremely destructive speed of modernity and postmodernity based on the generalized amnesia of fragmented subjects that try to establish bonds of identity with places. The wave of musealization that Huyssen (1994) highlights is in line with this tendency, since “it is a direct effect of modernization and not an event on its margins or outside it. It is not the sure sense of the traditions that mark the origin of museums, but their loss combined with a deep desire for (re-) construction”. In this respect, notions of cultural heritage and museums stem from an analogous perspective.

1.2. The Contributions of the ethnographic perspective

The ethnographic perspective is based on the discourse of groups and holders of cultural assets. Its contribution to the field of cultural heritage studies involves, through ethnography, seeking to produce narratives that capture interlocutors’ perceptions about the social and cultural environment, the climate, the description of the landscape, its history, among other aspects related to the cultural asset studied. At the same time, ethnographies concern specific dimensions of fieldwork and written production.

A fieldwork encounter establishes a particular relation with the interlocutors, a meeting of two worlds - one that is taken to the field by the researcher and another encountered therein. It is the essence of the anthropological / ethnographic experience, whose record privileges not only the researcher's view of the respondents, since the narrative must demonstrate the affectation caused by the relationships established at the time of the fieldwork. The moment of writing, which constitutes the time of reunion, opens space for the idiosyncrasies of the collectivity, for the particularities of the subjects, in order to expose not only the norm, but its deviation. From this trained look that dialogues between the general and the specific, a specific conceptual and methodological framework of anthropological analysis is designed, in a specific time of reflection on the experienced events.

The choices influenced by the research conducted around its ethnographic writing

---


are, therefore, very particular results of the engagement with the group studied and the museum. In this respect, the specificities of each ethnography reside in both the very specific engagements and interactions with the natives in the field, and in the re-elaboration of the experiences through writing. The empathy of the researcher with the researched collectivity must be central to the point of taking the cultural universe of its interlocutors as crucial, giving the measure of its importance and in order to emphasize the need for specific linguistic understanding.

Thus, it is not an extreme separation between the field of ethnography in which fieldwork is carried out and the specific work of the anthropological métier, since the notion of experience and theoretical-methodological basis, this creative tension between research and theory, seem rather to be inseparable faces of the same coin. Essentials are dissolved over the natives, thus promoting the historical dynamism of the research group, without freezing and creating suspensions in the time of the natives' actions. Ethnographic writing, however, is a distancing from lived reality. Nevertheless, so as not to create an abyss between the researcher and the natives when trying to maintain the “co-age” between the time of life and the time of writing. Following the words of Vincent Capranzano (1977), ethnography would be a symptom of the confrontation with otherness, the dialectical attempt to put an end to the encounter, thus removing oneself from the lives of the ethnographers.

Possibly, the first point of cleavage in the produced ethnographies that take into account not only the specific context of our interlocutors, but how they think, feel and produce their own knowledge about their culture, is the moment by which the encounter, the relationship, and the interaction of the researcher with the universe of the other becomes the fulcrum of the narrative.

From the encounter of two compared worlds, which produce idiosyncrasies, ethnography produces a relationship between its interlocutors, which includes the researcher himself and his subjectivity. Nevertheless, it is possible to affirm that every ethnography is an autobiographical record par excellence, since it carries in itself, still, an ethnographic moment of analysis of the contexts that cannot be undersized. Ethnography is a form of dialogue not only about the other, but with the other. The multiple vowels, the cosmovisions, appear in ethnographic experiences as a priority issue: it is not only talking about the other, giving voice to the other, or translating the native universe. It is rather an interlocution between the world of the other and my world, ethnography being a bridge between native theory and anthropological theory. In the rupture of the production of a generic voice on the studied collectivities, it is

---

intended to build symmetrical relations in the ethnographic experience, propitiating the interlocution between the world of the other and the world of the researcher\textsuperscript{11}.

The native categories begin to dialogue, therefore, in the binomial objectivity / subjectivity, since in addition to simple ethnographic data, they deal with concepts and theories understood in the terms of others. The bone of contention in this field of speaking with another, not only of the other, is the bridge between anthropological theory with its related fields, ethnography and museology being profitable fields. The displacement of the axis promotes the search for another balance between what is relevant and structuring for the ethnographed group. Polyphony goes beyond an experience; it is the genesis, the narrative model, the paradigm of knowledge construction. The force of speech is therefore in the encounter; the power of knowledge lies in the interlocution\textsuperscript{12}.

2. Main results and conclusions

The research and management of the collection evidences both the process of knowledge construction, which is achieved not only through its material collection, formed by a diverse set of mobile and immovable cultural assets, but also, and perhaps most importantly, by the set of experiences that conform a certain history and memory together with a community. In this process, the notion of social cohesion is relevant in the process of selection, categorization of cultural assets in conjunction with certain value regimes, in order to converge artifacts into consecrated assets by means of collectivist logic; cultural expressions, ways and methods of heritage assets craft; rituals in public performances; people in community representatives and spokespersons. It is also important to establish an interface with the interdisciplinary reflections between museums, heritage and collections, in an attempt to collaborate with the dialogical reflections between the material and the immaterial of culture.

The results of the research demonstrate that ethnography can promote a relevant counterpoint on the universalization of concepts, reinforcing the necessary reference of decolonization of methodologies. How do we create the fictions about the worlds of others, whereas the natives cannot express themselves through their own categories, concepts, and methods? As collaborative projects, they are very close to the natives' point of view, without, however, giving them, after all, the seal on knowledge, given that the researcher ends up appearing as a mediator and translator and, therefore, who interprets, cuts, edits the content formalized in the text as final product? Writing can approximate or distance / exoticize the other, since mechanisms of discursive potential, of play between contexts, are triggered.


Within different models of ethnography possible in museums, how do the native categories begin to dialogue, therefore, in the binomial objectivity / subjectivity, given that in addition to simple ethnographic data, they deal with concepts and theories understood in the terms of others? Therefore, in this field of speaking with another, not only on the other, the bridge between anthropological theory with its related fields, ethnography and museology having profitable fields. The displacement of the axis promotes the search for another balance between what is relevant and structuring for the studied group. The polyphony becomes beyond an experience, it is the genesis, the narrative model, the paradigm of knowledge construction.

This research bases its investigation beyond the analysis (identification, selection, clipping, description) of objects, but in relations with people from different social circles and their different sectors, agents and agencies, as a way of understanding this world and its relations between subject / object, the material / immaterial, the public / private. In such a way, the ethnography of a contemporary museum promotes the debate of its specific social practices related to the modality of museum associated with it.

References


Radia Bernaoui, H. Peter Ohly, Kebbab Salim, Mohamed Hassoun

The Mobile phone between challenge and expectations: a potential for information sharing between Algerian breeders and veterinarians

Abstract

Today, in addition to being informed about their sector, livestock actors want to be constantly connected to the outside world. In Algeria, among the various ICT supports, the mobile phone is currently the most accessible tool.

Does this revolutionary tool contribute to the reduction of animal losses in the livestock sector and can it be considered as a tool for economic development? The aim of this paper is to try to determine how the two actors (livestock breeders and veterinarians) use the mobile phone, and then consider whether it may be regarded as an indispensable tool for the health protection of livestock and, accordingly, to what extent it can contribute to economic development.

1. Introduction

We are interested in the relationship between two key actors – livestock breeders and veterinarians – and their impact on livestock protection and market development of animal breeding. A livestock animal cannot be productive without ongoing technical and sanitary care, which essentially involves the farmer. Then there is the veterinarian, who is responsible for following this production by providing care and wellbeing to the animal. Finally, there are the authorities in charge of preserving animal health and acting as confederates of the two actors.

Today, in addition to being informed about their sector, livestock actors want to be constantly connected to the outside world. In Algeria, among the various ICT supports, the mobile phone (a smartphone with Internet applications and not just a mobile phone) is currently the most accessible tool. It fits well within the socio-cultural traditions of the country in terms of communication since it is based on speech. Furthermore, it brings together writing, sound and image, thus allowing connection between different social and age groups in a knowledge sharing framework.

Moch (2012) notes that "communication is an art, that of formatting and conveying information”, while Moussoki (2005) states that "information is communication of knowledge; it is also communicated knowledge". Thus, the newly emerged professions based on market competition, must be reorganized by integrating the new function of knowledge management. Knowledge management capitalizes and transmits knowledge within institutions including the use of collaborative tools to share and disseminate knowledge.

From this perspective, Abdelkader Djeflat, professor of economics, noted that Algeria must be integrated into this new era of globalization (2008). It is in this light
that he cites the intensified use of information technologies and communication, the central place that innovation is increasingly occupying in competitiveness. He also points out that companies, challenged by new tools such as ICTs and called to be more economically intelligent, technologically aware, and strategically alert, in the face of emerging new professions created by the unseen force of competition, are obliged to reorganize themselves to incorporate the new knowledge management function.

Do these revolutionary technologies contribute to the reduction of animal losses in the livestock sector and can it be considered as a tool for economic development? Gado (2008) states that "the use of ICT, particularly mobile phones, has significantly improved the economic performance of African countries and even has transformed Africa".

Given that from the social point of view, livestock is known as an area where communication and the need for information have always been intense, we are interested in finding out whether the means and tools of communication that have been used until now are up to date and effective. What would be the most appropriate means of communication to maintain a permanent and continuous relationship between these two stakeholders? What would be the best means of communication between these actors and the health authorities?

Taking into account that in Algeria, the socio-cultural environment of the breeding world is based on oral communication, could the mobile phone be a tool of choice that can strengthen the relationship between breeders and veterinarians? In developed countries, the establishment of efficient communication systems has contributed significantly to the health preservation of their animal resources and to the modernization and improvement of the productivity of their farms. Anseur (2009) notes that "today, the information technologies associated with traditional forms of information transmission could offer adapted answers to the information needs of Algerian farmers". Therefore, since the traditional mode of communication in national livestock systems is facing this transformation, and knowing that in Algeria the mobile phone sector has grown exponentially in recent years, we try in a first step to answer the question of how the mobile phone is used by the two stakeholders in the livestock sector, the breeders and the veterinarians. In parallel, we try to understand whether the mobile phone can be considered as an indispensable tool for the health protection of farms and consequently contributes to economic development, knowing that losses due to diseases strongly impact the national economy.

This research is based on the conclusions of Anseur, who reveals that "any communication system can only be effective and efficient if it connects its users" and taking into account the study conducted by Pasquati (2011), which attests that higher mobile phone penetration rates lead to higher GDP growth rates.

Our problem aims to address, on the one hand, the relationship between the
establishment by the State of different mechanisms that allow the development of livestock, particularly the improvement of the animal health system and risk management and the expansion of ICT in Algeria, particularly the mobile phone, which is spreading in both urban and rural areas.

2. Methodology

We conducted a mayor survey in the most densely populated central region of Algeria, which includes 9 wilayas (governmental districts; Region 1: Algiers-Tipaza-Boumerdès, Region 2: Tizi Ouzou-Bouira-Béjaïa, Region 3: Blida-Medea-Ain Defla).

Questionnaires and interviews were focused on:

- The use of the mobile phone (smartphone) as a professional tool
- The level of ICT knowledge through a smartphone
- The indispensableness of the mobile phone for the health protection of livestock as a contribution to economic development

The survey of our study on the use of the mobile phone between the two actors, in this case breeders and veterinarians, was carried out between May and August 2016.

In order to give more credibility to our survey, we defined a minimum number of respondents, which we set at 300 for individual veterinarians from the private and the public sector (out of a total of 1,265) and 350 for individual breeders (in accordance with the relative number in each wilaya based on thorough statistics from the chamber of agriculture - DSA). The return rate was 92% and 90%.

As a sampling procedure we chose quota-sampling, specifically; sex, age, and region of practice, besides the two professional categories. For the pastoralist population, we were unable to apply the quota method. In fact, it was impossible for us to respect equity, especially in the "sex" variable, given that the population of male breeders is much larger.

As our study concerns two types of populations, we have developed two questionnaires with 50 easy understandable closed questions (some multiple-choice or dichotomous with open follow-ups): one for breeders and one for veterinarians. Apart from a few questions dealing with the professional field specific to each of the two populations, the two questionnaires are approximately similar, written in French. In the development of the questionnaires, we consulted several experts, such as information scientists, breeders, veterinarians and health technicians. Faced with the large number of questions, we imbedded the topics into the context of daily life. This approach facilitated an easy understanding of the questions and avoided a possible "stress" effect. Particularly among the population of breeders there are some who are sometimes very old and whose illiteracy rate is quite high. For this reason, we applied additional qualitative interviews by asking other simple and direct questions. During each visit,
which took place either on the breeding site and sometimes in the public areas frequented by breeders (cafés, small farms, etc.), we started, before submitting the questionnaire, with a thorough discussion. This allowed us to identify the communication problems facing breeders as well as the challenges that the mobile phone now offers.

3. Results: usefulness of the mobile phone in economic development

The results of our survey reveal that mobile phones and ICTs play an important role in economic development. Information and communication are fundamental for breeders, facilitating their work.

With respect to what breeders believe about the economic contribution of mobile phones:

- 35% it concerns access to information on the preservation of animal health.
- 21% it alerts the security services to cases of animal theft.
- 17% it is related to the sale of livestock products, the purchase and/or sale of animals and livestock equipment.

The beliefs of veterinarians with respect to the economic contribution of mobile phones underline the breeders’ opinions:

- 85% it is rather related to the fight against epizootics.
- 73% it contributes between 5 and 7 on a scale of 10.
- 47% it is an ordinary tool.
- 32% it is a tool to be developed.

With mobile phones breeders also receive better health care for their animals:

- 41% contact veterinarians between 5 and 10 times per week and 39% between 1 and 5 times per week.
- 82% the mobile phone actually contributes to improving the health care of animals.
- 73% the contribution of the mobile phone is between 5 and 7 on a scale of 10.

During our survey, we noticed that the size of farms has increased significantly, as well as the practice of raising livestock. The herds are increasingly being placed in modern buildings, substituting the traditional stables. As a result, both veterinarians and breeders will progressively have to use new technologies. The breeder will be able to carry out his various activities in a less laborious way, which will enable him to lower the need for labor force. Likewise, the veterinarian will be able to report "on the spot" all the data relating to health and production of the animals and to exploit them, again "on the spot", via his mobile phone.

The results of the survey show that a majority of veterinarians and farmers report disparities in the coverage of rural and peri-urban areas by mobile providers operating
in Algeria:

- 83% of veterinarians and 60% of breeders believe that there are disparities in the coverage of rural and peri-urban areas by mobile operators operating in Algeria.

We reveal that for the health and economic management of their farms, breeders need mobile phone applications for providers.

- 42% wish to have mobile applications indicating the addresses of agricultural institutions and bodies, e.g. health, bank, offices.
- 28% say that the best mobile application concerns animal traceability (electronic identification of livestock) via an easy electronic chip system.

4. Conclusions

4.1. The mobile phone: an efficient communication between breeders and veterinarians

Regarding relational and professional objectives, farmers are more involved in the management of health risks. Indeed, thanks to the mobile phone, they are increasingly contacting veterinarians, allowing them to strengthen these relationships, whose effects are as much social as economic. Thus, as demonstrated by our results, this relationship between the breeder and the veterinarian, which evolves positively thanks to the mobile phone, is no longer limited to the area of health but extends to the economic activity of the two actors as the mobile phone is able to reduce the vastness of the territory. The complex dimensions of the territory in several regions of Algeria, makes communication difficult. That a part of the national camel population was devastated in 2015 by an unfamiliar disease in the region of Bésbes, a remote village located 200 km southwest of Biskra, demonstrates the interest in using the mobile phone as a means of communication.

4.2. Mobile phone use as part of the breeding council

Moreover, the results of our survey clearly show that the use of cell phones by livestock farmers in a context strictly related to the health status of livestock is far from being an exceptional fact. In fact, the mobile phone is no longer considered by breeders only as an emergency tool but rather as an accompanying tool in the monitoring of livestock, starting with advice and health prevention - two functions that have an economic impact on livestock production and productivity. In general, we can conclude, with regard to the use of the mobile phone by breeders and veterinarians, that the tool in question strengthened the relationship that already existed between them but was limited to emergency.
4.3. Mobile phone, an economic development tool

The results imply that the use of the mobile phone by the two main actors of the livestock sector is felt economically. The use of the mobile phone has influenced the processes of communication and exchange of information between the actors of this productive sector, which is, moreover, of focal importance.

Our results show that, since the vast majority of farmers use their mobile phones for work purposes, more than half of the respondents consulted veterinarians via the mobile phone to inquire about the health status of their animals. In the same way, almost all veterinarians use the mobile phone for inter-professional communication, while most veterinarians receive messages from breeders. We therefore believe that the use of the mobile phone by livestock actors has greatly improved working conditions, just as the use of this tool contributes to better management of animal breeding as it overcomes travel limitations while also saving time.

In fact, our results show that we are witnessing a growth in the use of mobile phone technologies whose impact is both economic and social. Access to information (especially technical), via mobile phones, and ICTs in general, give the two actors in the livestock sector not only security, by limiting losses, but also increased efficiency giving them better production and productivity.

4.4. Livestock sector in a digital world

If the telephone is the communication tool of choice in the world of breeding, then the socio-cultural changes taking place in this world, which are mainly due to the expansion of ICTs, call for the implementation of the mobile internet in the agriculture sector. This survey has pinpointed the delay in the deployment of the mobile phone (including mobile internet) by mobile providers as well as by the government in the agricultural sector. Indeed, in a country as large as Algeria, the availability of a range of mobile applications and farm management software that directly connects the farmers to the veterinary office and/or the institution of animal health would certainly have a significant impact on the national economy. It is the same for the launch of warning systems against the various epizootics that repeatedly decimate the national herds.

Despite its results, which point to weakness on the part of mobile providers and application developers who are experiencing an enormous delay in terms of attendance to issues in the world of agriculture, it must be noted that major extensions to the communication network are currently being developed by mobile providers within the scope of total national coverage. The next step in the expansion of mobile telephony across the Algerian territory will be a fast Internet, which is a more profitable market compared to that of voice telephony.

This perspective can be realized in a very short time, especially given the diverse
range of opportunities such advances would bring: developed infrastructure, improved literacy skills of the breeders, democratization of mobile use, growing agricultural markets, et cetera.

References


Negotiating participatory KO in crowdsourcing infrastructures

Abstract
This paper investigates a participatory KO process in a crowdsourcing project where participants can make suggestions changing the KO schema of a historic demographic database. The KO process is conceptualized as participatory design of an information infrastructure. By identifying and analysing suggestions from the user community, implications on the database of participatory KO are described. The KO changes are also explained as related to interests of the different agents active in the project. The study shows how participatory KO can be used by cultural heritage institutions to balance the crowdsourcing project interests and perspectives of agents involved.

Introduction
Knowledge organization (KO) in archives, libraries and museums has traditionally been a domain of the professionals working in those institutions. However, participatory KO, introduced as a measure to tackle increasing demands from society for greater engagement of diverse user groups and their inclusion in cultural heritage and social memory, constitutes a new challenge for these institutions (Flinn 2010; Phillips 2014). This particularly applies to crowdsourcing projects, where new knowledge is produced by an external community of non-professionals but information is still expected to be organized according to institutional culture and traditions. To overcome ubiquitous institutional dominance and build a better digital working environment for the user community, which is indispensable for a successful crowdsourcing project, institutions have been recommended to “design [systems] in iterations” (Ridge 2012) and to tune the design to align with the needs and wants of their participants (Liew 2015). Those recommendations are in adherence with suggestions stemming from the context of participatory design (for example Star and Ruhleder 1996; Karasti and Syrjänen 2004). Their tenet is that inviting users of systems or information infrastructures (IS) in as co-designers creates a better system environment for all stakeholders, for example by taking into consideration the needs of different user groups.

To investigate how participants can be granted more influence in crowdsourcing, this study observes a project in which participants, besides transcribing handwritten funeral records to construct a historic demographic database, also submit suggestions on the design of the KO schema of this database. The studied community engagement for structuring the transcribed data is a KO process, conceptualized as participatory design of an information infrastructure. The analysis draws points from KO and IS theorizing to understand how suggestions are implemented and how they are changing
the results of the project. Hereby, conditions and relevance for participatory KO in cultural heritage crowdsourcing projects are explored.

Objectives

This paper investigates how participatory KO can be used to align the interests of different user groups in an IS for crowdsourcing. More specifically, the research questions for this paper are:

- How are the participants’ suggestions changing the KO schema for the database and how can these changes be explained by information interests of the actors in the project?

Machine-readable funeral records through crowdsourcing

The crowdsourcing project studied was initiated by Københavns Stadsarkiv (Copenhagen City Archives) with the aim of transcribing digitized funeral records from Copenhagen between the years of 1861 and 1912. The records contain information such as names, addresses, occupations, ages at the time of death and causes of death of the citizens of Copenhagen. Participants are allotted a more influential role than in many other transcription initiatives by being provided with an online user forum that allows them to discuss their work but also to present ideas for improvements to the project or the design of the database schema. Potentially, it gives participants the possibility of contributing beyond the mere task of transcription and thus more influence over the information they produce in the project. Implemented suggestions can reshape the database KO schema, affecting what information and search combinations it will be possible to extract from the database, which regulates its functionality and usability for different user groups. However, there are limits to user influence. All suggestions have to pass professional approval, meaning that the archive maintains control over all changes made in the database. This makes the user forum act as an arena for negotiations of KO principles between participants and professionals, especially in cases where the interests of those groups are divergent.

Professionals and participants in this study were both defined as agent groups. The professionals consisted of archivists working with the project at the Copenhagen City Archives. Participants were defined as the crowdsourcing community. They are partly comprised of amateur researchers and genealogists but also include anyone interested in contributing online to the transcripts. A third agent is academics, defined as professional researchers involved in projects at an academic institution and with a possible research interest in the database. Their claims influenced the KO negotiation indirectly even though they could not be observed as an active agent in the user forum.

While all of these agents had interests in the resulting database, they did not have equal influence on the design process of the KO schema. Academics were invited in early on in the crowdsourcing project and their information needs relating to their
current research projects were catered for in the initial project phase. Representatives from non-academic genealogy communities were also involved early on in the process but, as a more diverse group, they could not present equally well-defined objectives as the academics. Inviting participants as agents in the KO process was therefore essential to secure the project relevance of this group, in adherence to the rule that “good participatory design makes the problematics of use visible” (Neumann and Star 1996, 231).

**Participatory KO and ontologies**

As mentioned before, the IS studied in this article supports a metamorphosis of originally analogue, handwritten records to machine-readable information. Machine processability is attained by structuring the data in a database, but to ensure usefulness and validity of a database for all agents involved, an ontology is needed where mutual definitions and understandings of concepts are stated. Ontologies are based in shared understanding of the members of a domain or community (de Bruijn and Fensel 2010) and are commonly defined in information research as “agreed between a number of agents [leading to] effective communications of meaning, which in turn leads to other benefits such as inter-operability, reuse and sharing” (Agarwal 2005, 504). Infrastructures have been observed to work as experimental systems in moulding practical ontologies by joining new elements and generating relations between them (Jensen and Morita 2017). In different contexts – from small communities to the global level – ontologies play an important role as common language for information infrastructures. Correspondingly, ontologies can also be of varying significance. For example, the ontology created in this study (in the form of a database schema) is classified as a lightweight ontology.

In conclusion, an IS can serve to define the presence of concepts and terms and their use and relations within its domain through the construct of a lightweight ontology.

**Participatory design of crowdsourcing infrastructure**

All crowdsourcing initiatives need a base on which involved agents can meet, information can flow and commitments and community can grow. Based on how information infrastructure is defined by Star and Ruhleder (1996), a crowdsourcing system together with a user forum and digitized source documents can be characterized as an information infrastructure, a boundary object that makes communication between the agents possible. The KO schema of the database is part of this infrastructure which in the light of insights from studies of participatory design of information infrastructures makes it possible to discuss the participatory KO design of the database.

Among the eight characteristics of IS defined by Star and Ruhleder (1996), some of them are especially relevant in regard to the participatory design of the database analysed and its KO schema. Firstly, the criteria of the *installed base* on which the
crowdsourcing system rests includes the historical system of keeping count of citizens in Copenhagen and practiced in funeral records. It further links to the current installed bases of registers, finding aids and databases in the City Archives, as well as previous database projects. Installed bases of institutional classifications, standards and systems have earlier been shown to limit conditions for community influence and to influence the outcome of crowdsourcing after collecting user-generated information in archives and museums (Jansson 2017). Despite the installed base being such a critical element, participatorily designed infrastructures sometimes overlook it entirely, while on other occasions meticulously incorporate it in ways that would not have been possible with a less involving design process (Karasti 2014). In this case, the installed institutional base for knowledge organization becomes a constraint of participatory KO and can pose a risk for conflicts in KO negotiations between agents involved in crowdsourcing.

Another characteristic of IS is the transparency of a functioning IS, and the IS only becoming visible when it no longer works. A related concept is point of infrastructure characterized by the moments when the IS can no longer serve its purpose without being extended or modified (Pipek and Wulf 2009). This was selected as a theoretical concept to illustrate the provoking factor behind most modifications of the KO schema.

Furthermore, embodiment of standards is another IS quality and a central concept in this study as participants have to relate continuously to historical standards (like place names, spelling, or old-fashioned denominations of causes of death) as well as a plethora of contemporary standards connected to the installed base. Part of the participatory design process, therefore, is to reach a joint understanding, agreed to a reasonable degree and shared by all agents, of how to normalize the historical standards to fit a contemporary KO schema. The appearance of standards is thus a provocative factor for creating internal ontologies because it demands all agents to agree to using a mutual language to make cooperation possible.

Finally, the need for ontologies can be explained with object worlds, a concept that further enhances the understanding of the perspectives of the different agents in the participatory design of infrastructures. The concept was used by Neumann and Star (1996) after first being introduced by Bucciarelli (1988), not as a criterion of IS but as an element in IS design, symbolizing the special frames of thinking in different professional (or other) domains referring to methods, language, symbols, models etc. Archivists or academic researchers may, for example, be based within a different object world than genealogists. This implies that each agent involved in the IS had their own objectives for the functionality and use of the IS and that the interests of the agents was interpreted as based on their object worlds.
Method

This study is based on an analysis of a combination of several different source materials, the main part consisting of analysis of participants’ suggestions for KO modifications, posted in the user forum. This analysis is combined with analysis of regulatory documents for the project.

The forum discussions about KO suggestions were analysed to identify points of infrastructure (as defined by Pipek and Wulf 2009), i.e. where participants realized that they are in need of modifications to the database structure to continue their work in a satisfactory manner or simply to experience more contentment in their voluntary engagement. Every identified point of infrastructure resulted in a suggestion of KO and sometimes in a negotiation process between the agents.

The arguments were analysed in aspects of motivation and research interests for the identified agents. Arguments and interests of the user community were then compared to those of professionals and academics.

Changes to the KO schema in both complexity and breadth

The participants’ KO modifications can be categorized into four groups: New categories, Format, Classification and precision and Vocabulary extension. The meaning of the KO schema and how KO modifications answered to agents’ interests will be discussed below for each modification category.

New categories and ontology expansion

The participant influence on the KO schema of the database focused mainly on the inclusion of new data categories. In other words, user suggestions expanded the ontology by adding data types to the database that from the start had not been deemed important by either professionals or academics, although the information existed in the records. For example, “Place of death” and “Institution” (as an alternative to address of residence) were added. It was as if the participants took pride in performing as complete a rendering of the original records as possible. This addition of categories expanded the ontology of the crowdsourcing infrastructure with new elements. Transcripts were also supplemented with the serial number of the funeral records, to secure the connection with the database item and the original document. A field for free-text comments was also added for comments or comparisons with other source records, for example church registers. The comments field and serial number strengthened the connection to other genealogical source materials and infrastructural bonds of the database with installed bases in other databases and registers.

Formats

Discussions of formats for data categories were few but did occur in the forum. The best example is how the age of the deceased was changed from being noted in years
and months to being specified in years, months, days and hours. This was closer to how
the age at passing was specified in the original records. The first format had caused
trouble for several participants, transcribing records concerning infants. Format
modification is part of deepening agents’ understanding of usability of the project but
is also about analysis potential since some data can become invisible if put in the wrong
format. The discussions on the age of the deceased revolved around what kind of
analysis would it be possible to do with the data which, when rounded off to months,
lacked information about young infants’ ages of death.

A modification was suggested by participants, who early in the project saw the gap
in research potential, but the suggestion did not gain immediate approval from professionals. It was only after academic agents also defined this as one of their
information needs that the age format was changed. This can be seen as a sign that
agents with more interest in analysis also had more authority in format issues.

Classification and precision of ontological elements

Another category of KO modifications was how to perfect the understanding of
ontological elements. Foster children, for example, were discussed; were they to be
seen as a relation to a relative, a kind of civil status, or was it rather an occupation to
earn one’s living? Even if suggestions to reclassify a data category were never
implemented, the forum discussion enabled agents to come to a consensus on category
definitions, despite their different object worlds. It thus contributed to a common view
of elements within the ontology, making the IS more stable and valid.

Expansion and classification of controlled vocabularies

The community was also active in expanding lists of controlled vocabulary of
institutions, cemeteries, causes of death, etc. Discussions about causes of death were
occupied with formation of standard forms, such as which of all the different mentions
of, for example, diphtheria, were to be selected as the main form for the vocabulary.
They also concerned how to classify different descriptions of the same disease, or how
similar two causes of death had to be to be classified as the same. This work constantly
included references to and evaluation of historical medical standards but also meant
much for participants’ interests and satisfaction with the project because of their
perspectives on historical conditions about life and death.

The relevance of participatory KO in crowdsourcing

Participatory design by a crowdsourcing community can expand the community
ontology in aspects of both quantity and quality. In the forum, the object worlds
sometimes collided, which gave rise to new discussions and brought the ontology of
the IS forward. The participatory KO process provided an outlet for bringing forth the
diverse arguments of the agents and created a more stable IS in the end, giving room
for the object worlds of the agents involved and making the IS relevant for a larger number of potential users. Participatory design of KO (although institutionally controlled) is thus bringing advantages to crowdsourcing ISs by reflecting and joining the object worlds of the agents in the emergence of a mutually agreed ontology.

Besides developing ontology, participants were also augmenting the original KO by adding links to the installed base of the database, for instance by insisting on connections to other registers and projects and by suggesting new fields for serial numbers and commentaries, but also when classifying and extending the understanding of ontological elements and concepts. This shows that even though bases of institutional KO can limit community participation, there are ways to work around them by adding more flexibility in KO systems and making participants part of the development process.

The arguments for KO modifications can partly be explained by the object worlds of the agents involved. This is exemplified in academics’ requests for data adapted for big data analysis (for example concerning registration of the age of the deceased) or in controlled vocabularies. In arguments deriving from the non-academic community, participants were more interested in documenting individual cases and aspects of uniqueness. This was manifested in interests connected to individuals, where every documented detail was of great importance as a part of the detective work of genealogy. They were also more interested in selecting out, for example, specific diseases or causes of death, rather than grouping them together. This attraction to detail and peculiarity can maybe be seen as a manifestation of the joy, pride and motivation to engage in and commit to the project. Participants were allowed to discuss and incorporate a part of their object world in the database. The professional agents, on the other hand, rejected suggestions that were too detailed to avoid causing unruliness in the database schema. They wanted to keep the ontology as simple and easy to grasp as possible, so as to not scare off participants from the project. This illustrates one of the complications of participatory KO as professionals have to unite discussions and negotiations between project agents with their professional responsibility for management of the project.

To conclude, participatory design of KO in crowdsourcing is therefore one possible method to resist “the standardization of a single, powerful group’s agenda” (Star and Ruhleder 1996, 114) and to take steps toward more democratic interoperability, reuse and sharing, echoing Agarwal’s description of the advantages of ontologies. In light of the present findings, user communities in crowdsourcing projects should therefore be engaged as co-designers of KO when possible, and invited to have greater influence over information infrastructures to which they themselves are contributing. This can improve the opportunities of a crowdsourcing project to be regarded as successful in the eyes of all its participants and not only those with direct influence over the institutional KO structures.
References


New ways to produce shared knowledge to improve cooperation in overcoming societal challenges in healthcare: the lever of innovative interface organizations in France

Abstract
Nowadays new digital devices have raised questions in the healthcare sector. They have given new possibilities to both improve care and master costs but may also create new constraints. In France the problem of “walling off” areas of healthcare, especially between primary care and the hospital sector, has led to the development of new healthcare interface organizations (HIO). They constitute innovative areas to develop coordination and cooperation between all the involved actors, including patients and their families. Using an information and communication approach, we analyze how HIO may represent an interesting lever of improvement through production of shared and collective knowledge both regarding the healthcare quality of empowered patients but also for the evaluation of the services proposed by these new HIO.

Introduction: context and objectives
Healthcare systems are in crisis in all developed countries with problems of costs and quality of care. The French healthcare system may be considered “out of breath” (Isaac 2014) being overly curative. ICT may help to transform it into a more preventive system within an e-Health context with new information, communication and shared knowledge challenges.

At present, France is tackling a specific problem: the “walling off” or compartmentalization of medical activities, especially between primary care and hospitals, is fueling social and territorial inequalities in healthcare, at both individual and collective levels. The French healthcare system is reasonably effective, particularly in term of access for all its citizens, but it is also very expensive – 12% of GNP – and thus inefficient.

Transforming a predominantly curative system into a more preventive one is a great challenge in terms of the new approaches required, from coordination to cooperation, in turning to new uses of shared data. Moreover, the key challenge is sharing data to build contributory knowledge favoring new cooperation between all the actors involved, including the empowerment of patients (cf. 2002 Law and the idea of “sanitary democracy”). All this in a context of “augmented human”, “quantified self” and full use of data (big and open) and their specific characteristics within the context of healthcare (regarding particular legislation), leading to the development of new socio-technical devices, especially platforms.

1. Researcher’s position and methodology
This paper was developed in the context of a Master’s degree and is based on
cooperation with health insurance offices and especially with new health interface
or ganizations (HIO) from an economic intelligence perspective of organizational
telligence (Wilensky 1967). The research is aimed not only at boosting the
competitiveness of companies but also building social cohesion in areas (Carayon
2003) faced with the particular challenges of e-health and inequality.

In the field of information and communication science, we work in accordance with
the perspective proposed by F. Bernard (2006) which sees the convergence of four
issues – connection (interaction), meaning, knowledge and action – in an “engaging”
framework based on building knowledge for (research) action, by insisting on the
complementarity of information and communication approaches.

We focus on A. Mucchielli’s (2010) “situational and interactionist semiotic”
approach with the importance of interactions in specific contexts to give meaning to the
actions of all the actors, including patients and their families. We also attach importance
to socio-technical devices, that is, ICT tools. We propose a position that we call ICOE
(Information and Communication for Organizing Ecosystems) which includes (public
or private) organizations, socio-technical devices such as platforms and of course,
territories. In accordance with the "organizing" dimension (Weick 1969), we advocate
the integration of socio-technical devices (as platforms of activities) in the wake of
Simondon (Du mode d’existence des objets techniques, 1958) and also Callon-Latour’s

We also look to the "Montreal school" (Cooren and Robichaud 2011) emphasizing
the importance of interactions and speech acts following the lead of research conducted
in France – “Language and Work” or Langage et Travail (Borzeix and Fraenkel 2005)
– to better understand the “iceberg of activity”, considering the importance of language
at work (communication challenges) to improve coordination with the aim of trying to
produce shared knowledge to build cooperation between all actors.

We extend this approach to the complementarity of communication (interactions,
discourse at work, etc.) and information: importance of data, particularly to build shared
knowledge for action, focusing on information and communication scenarios to
produce shared knowledge.

Following the lead of Le Cardinal et al. (2001), we also assert the importance of
trust, especially in the management of complex projects. In this respect, we highlight
its importance in regard to issues of recognition at work (self-esteem of actors) and of
professional identities, thus including affective aspects and representations of the
different actors.

As well as giving special consideration to innovation in relation to “everyday
activities” (Alter 2005) we also affirm the need for new intermediation in knowledge
production and, therefore, the societal dimension of innovation in organizational areas
considered as ecosystems (Godet et al. 2010).
According to M. Doueihi (2011) we insist on the role of social platforms “not because they manage the access and the storage of data, but because they have become thanks to the activities of the (human) users, areas of convergence between information, communication, knowledge and sociability” with all the importance of social networks and new digital identities.

This is all done with the intention of trying to improve cooperation in healthcare through new shared knowledge with the idea of constituting “strategic communities of knowledge” such as those described by N. Moinet (2009), which are of particular relevance to healthcare, with the idea that “collective intelligence is first the “intelligence of link” (Zara 2008). The challenge is then to develop conditions to promote collective intelligence rather than the individual intelligence of each actor. This objective meets the imperative need to tackle the great challenge of social and territorial inequality in healthcare (both individual for people and collective for territories) and thus, especially, the issue of “medical deserts”.

2. Healthcare Interface Organizations (HIO) as innovative areas between primary care and hospitals

New healthcare interface organizations (HIO) constitute innovative areas developed since the 1980s to reduce the tendency to “wall off” and create divisions between primary and hospital care, and so to promote cooperation between these two key areas of medical provision in France. First to appear were the Hospitalization at Home (HAD: Hospitalisation à Domicile) and Healthcare Networks (réseaux de santé), then Medical Houses (MSP: maisons de santé pluriprofessions), followed ten years ago by MAIA, for coordination of devices for Alzheimer patients and, more recently, PAERPA, for elderly people with little autonomy. We can also refer to CLIC, Information and Coordination Local Centres (centres locaux d’information et de coordination).

All focus on the challenges of coordination (Bloch and Hénaut 2014) to improve follow-up patient care, particularly through PPS or personal health plans (Plans Personnalisés de Soins) or patients pathways (parcours de soins). With the exception of MSP (which corresponds to a grouped medical exercise and not to a functioning network), these different interface organizations share the key mission of coordinating healthcare, not delivering it directly but organizing it. They are not new care organizations but contribute to improving the coordination of existing organizations.

They meet important challenges, the first being to show that their existence matters and that their services add value in a healthcare world often characterized by the individualism of its actors. There is also the risk of hard competition in same territories creating new divisions. Bloch and Hénaut (2014) insist on the visibility of their action in an integrative approach within a territory.

As already underlined, these health interface organizations, mainly created for
improving coordination of GPs in primary care, strongly rely on the development of new socio-technical digital devices such as electronic health records. In France DMP (dossier medical partagé) are increasingly being incorporated into platforms such as PTA (Territorial Platforms to Sustain Primary care Activities – Plateformes Territoriales d’Appui), experimentally developed with the TSN (Digital Health Territories – Territoires de Soins Numériques) (2013, 17), probably for integration in regional information systems. Information and digital devices are essential, and there are often digital as well as medical “deserts”.

The question of the services provided by these new tools is essential. It explains for example the success of Diraya medical records in Andalusia (Spain) and the failure of the DMP in France launched in 2004 with a lot of attention. We may also point out the importance of telemedicine activities especially for disadvantaged areas such as medical deserts.

Another key point is the notion of patient pathways (traceable and interoperable) promoted by these interface organizations, especially healthcare networks. But these are health organizations and not just cure organizations. Most patient follow-up care, often in complex situations (patients suffering from multi-pathologies, isolated and with limited financial resources) “is not mainly to cure” (G. Mick¹); helping GPs better manage these complex cases is the main justification for HIO.

These HIO must also justify the use of public money with regard to their evaluation. In France, evaluation is a large part of so-called “new public management policy” or NMP (Nouveau Management Public) for rationalization of public power actions. But there have been much abuse and suffering, with Ogien (2009) speaking of “the hospital seized by quantification”, and Gaulejac (2005) referring to “a managerial ideology resulting in social harassment”.

Mintzberg (2001) pointed out the excesses of the word “efficiency” (which has become a "bad name"), given the excesses of some quality approaches, which have generally become rigid quantitative controls that can lead to sanctions and not to development aid.

In this context of often difficult relations between interface organizations and their sponsors, we have worked on new approaches of evaluation, corresponding to an information and communication approach, based on the creation of new shared knowledge to improve services to patients while also involving them more.

3. From information and communication challenges to shared knowledge building

The initial investment, as already mentioned, may create a favorable context for sharing knowledge, especially with “strategic knowledge communities” (Moinet 2009),

¹ Gérard Mick, specialized doctor and researcher, is President of “UNRS (National Union of Healthcare Networks – Union Nationale des Réseaux de Santé).
but also with a strong dimension of knowledge production for action in everyday situations (better follow-up of patients i.e. traceability of pathways, to improve the quality of care). The aim is to promote a new cooperative spirit supported by new socio-technical devices and emphasizing the training dimension. These innovative interface organizations are two-way learning organizations: they train their members to spread the spirit of cooperation and openness towards other professionals and they also build a collective intelligence from the knowledge of all their members (professionals, but also patients and their families).

Some patients change considerably (Mick) and are more demanding, often becoming experts in their disease and actors in their own health, fulfilling notions of "health democracy" and "empowerment". The High Authority for Health (HAS – Haute Autorité de Santé) has proposed adopting certain process approaches in quality management including enhancing the role of patients: "tracer patients" (patients traceurs) and "expert patients" (patients experts) would put into practice the concepts of PPS, Personalized Project of Care (Projets Personnalisés de Soins), and care pathways.

In our ICOE position (see above), we insist on the "performatve capacity of information". "Information is an action, in the sense that it makes a representation" (Vigouroux and Zugasti 2017).

In partnership with healthcare interface organizations, we are working to develop innovative, qualitative, participatory and immersive methodologies to co-build shared knowledge devices with all actors, integrating their feelings and emotions and their development. We try to achieve this by converging representations and firstly by trying to develop a shared language around shared objectives to try to overcome the compartmentalization of different approaches to various jobs. It is also a question of apprehending and valuing the creativity dimension to favor cooperation in a dynamic of "improvisation and working together" (Ménissier, Martin-Juchat and Lépine 2017).

The new approaches centered on the information and communication issues that we propose also include the promotion of HIO by trusting in their capacity for continuous self-improvement. Here again, cooperation (communication) is built on new uses of shared data, more often produced collectively for shared purposes: traceability of care and valorization of the services provided to both patients and professional members of these interface organizations.

We are working on new approaches to care pathways around the notion of “tracer patients” with a project to create an observatory of care pathways. This will involve trying to converge three types of process: compliance with benchmarks, implementing change and development of innovation with improvisation and creativity (Caliste and Bourret 2015). Promoting adaptability is therefore essential.
4. New cooperative approaches to platforms and socio-technical devices for shared knowledge

All these perspectives aim to promote the appropriation and better use of new socio-technical devices (computerized patient records, platforms, telemedicine tools, etc.).

Computerized patient records constitute a first step, for monitoring patient care pathways. We compared the French DMP (Personal Medical File – *Dossier Médical Personnel*) and Diraya medical records in Andalusia (Spain). The French DMP launched in 2004 was largely a failure because of the defective project management behind development of this tool. There were several changes of operators and it was not a cooperative project with users, hence their fear of it being above all a control tool to boost the activity and revenues of Health Insurance or CNAMTS\(^2\) (French doctors are paid for each medical act). Since 2017, the DMP (having changed the designation for medical records belonging to patients from “personal” to “shared” records) has been entrusted to CNAMTS, which is experimenting with new approaches, making it one of the axes for improvement of the services provided to its users. The case of Diraya medical records is different, due to better project management: cooperation with users, continuity of the project team and development of real services to users (making appointments, health data management, etc.). It is a concerted device that has really focused on the development of shared knowledge; the word *diraya* means knowledge in Arabic.

Coordination of the different actors has become a key issue in France. The public authorities rely heavily on the development of intermediation platforms: both for general services for all citizens (cf. Ameli .fr platform for CNAMTS) but also specific services to improve cooperation for primary care professionals. This is the case of PTAs (Territorial Support Platforms – *Plateformes Territoriales d’Appui*) for primary care and, more recently, SNACs (Digital Services for Coordination Support or *Services Numériques d’Appui à la Coordination*) emphasizing both notions of care pathways and new areas of care (*territoires de soins numériques*\(^3\)).

As with any technique (Ellul 1990), these new tools are ambivalent. They can both contribute to improvements but also, by creating new constraints, as ERP\(^4\) they risk causing discouragement (Mayère 2017) owing to new forms of “tailored services” or “uberization”.

There are many tensions between costs, standards and values among the different job cultures that need to be taken into account. The rigid models often proposed by National Healthcare Offices frequently lead to despondency and suffering at work and rarely meet the needs and the particular context of the different interface organizations

---

\(^2\) CNAMTS: *Caisse Nationale d’Assurance Maladie des Travailleurs Salariés.*


\(^4\) ERP: Enterprise Resource Planning, in French PGI (*Progiciel de Gestion Intégré*).
and the specificities of their (complex) patients or their professional actors.

A customization approach defining invariants for all structures but with a large part left to local adaptations could be an interesting way forward. It is a matter of developing a collective intelligence (contextualized for each interface organization) in a perspective of continuous improvement of the service provided to the two types of users (professionals and patients) by also integrating the other actors (public services, local authorities, etc.).

One essential aspect is to insist on incorporating the indispensable human dimension into these socio-technical devices, in a context of co-production of services through the cooperation of all actors. Accordingly, ethical issues are unavoidable and "technical performance must be up to human requirements" (Thiel 2003).

**Conclusion**

Health Interface Organizations (HIO) thus constitute spaces for innovation and the development of collective intelligence to develop and improve patient services, which are interesting to study in the context of the profound changes provoked by e-health devices.

The production of shared data leading to a form of collective intelligence can greatly contribute to the improvement of the services provided to the different users of these interface organizations (health professionals but also patients and their families) and bolster their legitimacy for sponsors.

The first step, often through training activities, is to converge the representations of all actors to create the conditions for collective intelligence development in a perspective of continuous improvement. In a second step, the appropriation and adaptation (customization) of the socio-technical devices proposed to these interface organizations must make it possible to produce shared knowledge and then collectively set up new modalities for more contributive evaluation methods in a context of continuous improvement developed with involvement of all the actors.

New health interface organizations (HIO), involving public and private organizations and fostering cooperation among all the actors, including patients and their family, can play an essential role as spaces for innovation, building cooperation around shared knowledge. If healthcare is still a competence of national states, the opportunities offered by cooperation at European level should not be forgotten, such as the Epos (Smart Open Services for European Patients) project for patient mobility in the European Union. Based on shared knowledge, these health interface organizations (HIO) therefore constitute an important lever of personal mobility and integration in the European area.

---

References


António José de Bastos Leite, Francisco Carlos Paletta, Maria Fernanda da Silva Martins, Teresa Silveira

The Role of Neuroscience in information and knowledge appropriation

Abstract
Understanding how knowledge should be organized, requires for information & library science to embrace brain science as a full partner. This brings with it a new perspective from information behavior; on the basis that behavior should constitute reading skills, memory and thinking. Currently, the reading behavior of “Digital Natives” versus “Digital Immigrants” has been the object of attention in several areas of knowledge, such as sociology, psycho sociology, and information science. Brain science, together with imaging, have also approached the problem of reading, connecting it with memory and thinking. This research deals both with reading and information science, as well as neuroscience, neuroimaging and cognitive psychology. The aim of this research is to differentiate the areas of the brain that are activated in individuals belonging to the generations born in the digital context and in those who have adapted to it. In terms of methodology, the study participants consisted of 24 right-handed male individuals, 12 of whom, aged between 19-30 years old, comprised the subgroup of “digital natives”, while the other 12, aged 42-50 years old, served as “digital immigrants”. A condition involving sequential reading (TpA), and a condition involving digital reading (TpB) were used. At the end of the fMRI scanning, participants were asked a set of questions, to assess their degree of understanding, retention and capacity to use information from the fMRI paradigm. The results did not reveal major differences in global terms. However, significant differences were detected between the two conditions of the paradigm, by contrast TpB - TpA, in the right fusiform gyrus in both groups. The peak of statistical significance occurred at coordinates 33 -52 -14. The questionnaire revealed differences between groups. We conclude that differences found in the results of the questionnaire could be related to the effect of distractibility on the retention of information when other stimuli are shown concomitantly to the sequential reading. Certainly, future cognitive tests and imaging work focused on the processes underlying the attention, memory and use of information should be done to clarify this. In the meantime, information professionals need to consider information according to the binomial “push and pull”; that is, products and services should deliver answers to the users but at the same time should encourage them to get more, to explore different perspectives. Increasing curiosity is the key to personal knowledge development. Finally, information products and services should avoid the current information trend, in some way expressed in this research, of get in, get the answer and get out. This leads us to the main challenge, namely that organizing knowledge reinforces the focus on user knowledge and the need for brain science and physiology to secure thinker users instead of information reactive ones.

Introduction
In information science, the study of informational behavior involves the consideration of the effect that lived experience has over the individual (one's mind), as well as the impact that the environment has in the functional development of one's brain. This latter concern seems to be more implicit, perhaps because it demands the establishment of scientific connections between two fields which traditionally have had difficulty in dialoguing: exact or life sciences and social sciences. Hence, one of the challenges of this research: sustainable interdisciplinarity.
Accordingly, to describe a technological experiment conducted in this scientific domain, our attention must first be drawn to two key concepts: digital inclusion and digital literacy. Whereas the former is widely associated with operational issues in the domain of equipment, the latter invites complexity into our reflections, since it adds to the operational domain both an informational domain and a reader. In other words, to evaluate the attitudes of the reader, how he uses information and develops knowledge, involves understanding how he retains this information. In order to do this, it is necessary to investigate how a subject reads and how he processes and retains what he has read. In practical terms, this means determining the level of reading competence and the brain areas which are activated during reading, relating this information to the attitude the reader showed in the ways he uses and develops personal knowledge. Thus, this study aims to verify to what extent the new formats of reading influence the retaining and use of information in individuals who clearly grew up in the context of technological mediation. In practical terms, we intend to:

- Differentiate the brain areas which are activated during the reading of static formats (text) and dynamic formats (animated text and image), both for subjects who were born in an age of early contact with digital formats and those who had to adapt to them at a later age.
- Relate this information to the informational behavior of these subjects (how they retain and use information) using the results obtained, relating to the future of information users, of information services and products, as well as of informational professionals.

An interdisciplinary approach to informational behavior

Whereas biological sciences describe in a single manner the way the brain activates different capacities and transforms them into competences, social and human sciences offer multiple designations for the generation born after the 70s/80s. In the literature, the description of the processes that activate the brain to become a reader refers consistently to the same concepts: use during the sensitive periods (especially those which occur between birth and 3 years of age, and between 6 and 12 years of age), the workings of memory, both repetitive and elaborative (Wolfe 2004), the need to develop/exercise memory and the control of selective attention. The same consistency is not found in the literature when describing cognitive functions. The term “digital native” is one among many such designations, others including “new millennium learners” (Pedró 2006), “net generation” (Tapscott 1999), “gamer generation” (Carstens and Beck 2005), “generation Y”, “generation M” (media), “generation V” (virtual), or even “generation C” (Veen and Vrakking 2006; Rideout et al. 2005), in which “C”, according to the authors, refers to three behaviors/functions characteristic of this generation: connectivity, creativity and click. This plurality of terms for those who were born in a digital environment is not found in the literature in reference to the
previous generation. The most common designation is “digital immigrants”. However, taking into consideration the abundance of labels for the former group, we could ask ourselves if this amounts to different representations of this generation.

**Methodology**

**Participants**

Our sample was composed of 24 right-handed male subjects, without any neurological alterations, who did not take any medication capable of influencing cognition. These 24 subjects made up two subgroups: 12 “digital natives” between the ages of 19 and 30, and 12 “digital immigrants” between the ages of 42 and 50.

The choice of a sample composed entirely of male subjects was made as a result of previous studies (Silveira 2011) which showed that male participants had a significantly more homogeneous reading behavior than female participants.

The age groups selected correspond to generations which were born and raised in distinct technological contexts (i.e., before and after the digital environment). No participants were chosen above the age of 50 in order to prevent possible biases resulting from the aging process. Likewise, no participants were chosen below the age of 18 so as to guarantee full reading capacity and competence.

To be selected as volunteers and take part in the study, the candidates had to fulfill the following conditions: (a) master the required reading competence, that is, decode, understand, and interpret written messages; (b) have a minimum reading speed; (c) have or be in the process of obtaining an undergraduate degree at university; (d) differentiation regarding the implicit educational contexts.

To ensure the fulfillment of these prerequisites, especially (a) and (b), a number of activities were carried out prior to the realization of functional magnetic resonance imaging (fMRI) studies and the application of post-fMRI questionnaires.

All the candidates were subjected to: (1) a questionnaire for the characterization of the sample, so that we could get to know and understand reading habits, learning contexts, and the development of reading taste; (2) an assessment of reading competence, with the purpose of ensuring that none of the volunteers were purely functional readers; in other words, candidates should have been able to interpret the texts perfectly; (3) an assessment of reading speed, with the purpose of warranting behavioral homogeneity with regard to reading speed, as well as the adequacy of this reading speed according to the paradigm of functional magnetic resonance imaging. Candidates should have been able in 40 seconds to read a block of text with characteristics similar to those of the text used in the study.
Instruments

The research instruments used in the study were: a semi-closed questionnaire, previously to the fMRI; a functional magnetic resonance imaging study, undertaken during the progression of the block structure detailed below; and a closed questionnaire, after the fMRI (Quivy 1998).

A – Functional magnetic resonance imaging protocol

Images were obtained using a magnetic resonance equipment model 3 Tesla (Magnetom Trio, A Tim System, Siemens, Erlangen, Germany) equipped with a 12-channel antenna.

An fMRI sequence was obtained based on the BOLD signal with the following characteristics: echo time [TE] = 35 ms, repetition time [TR] = 3000 ms, flip angle = 90º, field of view [FOV] = 192 mm, section thickness = 2.5 mm, no interval between sections, number of repetitions = 200, acquisition matrix = 92x92, spatial resolution = 2.5x2.5x3 mm, acquisition time = 10 minutes.

A T1-weighed, high resolution sequence was also obtained with the following characteristics: TE = 3 ms, TR = 2300 ms, flip angle = 9º, inversion time = 900 ms, FOV = 240 mm, section thickness = 1.2 mm, number of sections = 160, acquisition matrix = 256x256, spatial resolution = 1x1x1.2 mm, acquisition time = 9:14 minutes.

B – fMRI data collection

A block structure was used alternating eight periods of rest with seven blocks of activity, consisting on the reading of an excerpt from Lewis Carroll's Alice's Adventures in Wonderland (cf. Figure 1 in the annex).

The length of each block was 40 seconds. The periods of activity were subdivided in two conditions: sequential reading, without the interference of animations or additions over the written text (TpA); and reading of the same text with the interference of animations (i.e., illustrative images and movements). Figure 4 represents the block structure. At the end of the fMRI session, the participants were asked to answer a set of questions.

C – Image processing and analysis

We used the Statistical Parametric Mapping (SPM) software for spatial processing, which included the realignment of functional images, the segmentation of the weighed structural images into T1, the co-registration between structural and functional images, as well as the normalization of functional and structural images for a template standard with a 1x1x1 mm spatial resolution. We also applied a smoothing algorithm to functional images, with full-width at half maximum Gaussian kernel of 6x6x6 mm.

FMRI results were obtained through a standard SPM analysis of global brain activity. This type of analysis used the general linear model based on the codified
stimuli in the visualization conditions. Movement parameters based on the realignment procedure and high-pass filters at 128 seconds were also used.

D – Post-fMRI session questionnaire

This questionnaire, the objective of which was to assess the degree of comprehension of the fMRI paradigm and test the levels of retention and use of information, was composed of three closed-answer questions. These questions, with differing levels of difficulty, aimed at evaluating the ability to understand direct textual messages, and to retain the details and the subtlety of the written message. This instrument was also intended to verify the impact that animations in the reading may have on the processing/retention of textual information.

Each of the three questions offered five options of answer, among which the volunteers had to select only one. Additionally, the participants were asked to make a drawing about the message they had read, which was expected to represent faithfully that message. Inconsistencies in terms of omission or addition in relation to the content shown were registered and evaluated under a nominal variable.

Results

Functional magnetic resonance imaging

These results showed no significant differences in global terms. Specifically, no statistically significant differences were detected between the groups of “digital natives” and “digital immigrants”, as shown in Figure 2 in the annex.

However, significant differences were detected between the two conditions in the paradigm, in the contrast between TpB and TpA, in the right fusiform circumvolution, as shown in Figure 3. The peak of statistical significance occurred at the coordinates 33 – 52 – 14 (marked in red in the table of Figure 3), in a “voxel” of the right fusiform circumvolution, located in part of the right temporal and occipital lobes.

Questionnaire

Taking into consideration the questionnaire, a statistically significant difference was found. It occurred in question 2 (Q.2), table 1, which was designed to test the capacity of detecting and retaining text minutiae, differently from what was intended by questions 1 and 3. In these questions, no statistically significant difference was found, as it can be seen in tables 2 and 3 below.

Table 1: Retention of unread information - Q.2

<table>
<thead>
<tr>
<th></th>
<th>Frequency (yes)</th>
<th>Frequency (no)</th>
<th>$\chi^2_{(1)}$</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natives</td>
<td>1</td>
<td>11</td>
<td>6,750</td>
<td>0.009</td>
</tr>
<tr>
<td>Immigrants</td>
<td>7</td>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2: Retention of unread information - Q.1

<table>
<thead>
<tr>
<th></th>
<th>Frequency (yes)</th>
<th>Frequency (no)</th>
<th>$\chi^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natives</td>
<td>8</td>
<td>4</td>
<td>0,202</td>
<td>0,653</td>
</tr>
<tr>
<td>Immigrants</td>
<td>9</td>
<td>3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Retention of unread information - Q.3

<table>
<thead>
<tr>
<th></th>
<th>Frequency (yes)</th>
<th>Frequency (no)</th>
<th>$\chi^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natives</td>
<td>8</td>
<td>4</td>
<td>0,889</td>
<td>0,346</td>
</tr>
<tr>
<td>Immigrants</td>
<td>10</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As far as information use is concerned, statistically significant differences were found when the two groups were compared. Even though the issue we address below was not originally intended to be analyzed in this study, it seemed interesting to show the results we obtained when the subjects were asked to make a drawing illustrating what they had read. A statistically significant difference was found between the groups, as shown in table 4 below.

Table 4: Use of information

<table>
<thead>
<tr>
<th></th>
<th>Frequency (yes)</th>
<th>Frequency (no)</th>
<th>$\chi^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natives</td>
<td>1</td>
<td>11</td>
<td>16,667</td>
<td>0,000</td>
</tr>
<tr>
<td>Immigrants</td>
<td>11</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the execution of this task, “digital natives” almost always added to the central character – Alice – a small number of elements about/relative to the action of the character, not seen during the animated reading period. On the other hand, “digital immigrants”, when facing/executing the same task, no more than/only replicate the visual elements seen during the period, without adding to it any other details.

Discussion

The results show that there are similarities and differences between the groups in the study. The fMRI study showed that “digital natives” and “digital immigrants” do not activate different brain structures when these subjects read a text. Nevertheless, animated-format reading differs from sequential reading regarding the activation of the right fusiform circumvolution. This area has very important functions, since it is involved in facial recognition, color information processing, as well as in the recognition of words and in the identification of categories of variables (Mascaro 2008). Additionally, it is implicated in a wide network of attention processing: when we focus
on a point or an object in space, there is an increase in the blood irrigation and in the electric activity in the fusiform circumvolution (Mascaro 2008).

Attention plays a very relevant role in different brain functions, since there is the engagement of brain structures and processing networks related to sight, to hearing, to touch, and to the memory, which, collectively, allow for the construction of what we learn to recognize (Mascaro 2008). Thus, the fact that the activation of the right fusiform circumvolution does not occur in the same circumstances in the two reading conditions of our study, in either of the groups, might mean that the introduction of other elements, in addition to the written text, could be perceived as an “add” and not as part of the text per se.

A study published in 2009 by Gary Small et al. suggested that internet search – which requires the combination of text and images – when focused on a specific objective, even if subjected to different types of stimulation, improves cognitive performance. In this study, individuals with excellent capacity and competence in reading and navigating the internet activated more brain areas than those who did not possess these capacities/competences (vide Activations for the book Text reading and internet searching task in comparison with the baseline nontext bar task).

In the same direction, the present results seem to indicate that the multi-stimulation of the written text may improve cognitive functions and the working of brain structures, as long as reading capacity and competence are automated processes. However, permanent exposure to technologies and their use, especially in a situation of precocious overstimulation, when the brain is in very malleable stages, could be harmful.

This possibility encourages us to engage in a future investigation about this issue, namely, examining the consequences of overstimulation in the filtering, processing and use of information in individuals who modeled their brain structure and function to respond to multiple simultaneous stimuli.

Although this is not intuitive, the absence of demonstrable differences in the “reading brain” between the two groups does not invalidate Interlandi’s words: the brain’s flexibility (plasticity) to adapt and change its output is based on new and reinforced informational stimuli or inputs (2008). In other words, even though “digital natives” are subjected to new stimuli and inputs more often (and earlier in their lives) than “digital immigrants”, the recruitment of brain systems involved in reading can be identical and, it should allow (in optimal conditions) both “natives” and “immigrants” a truthful and flexible interpretation of the world.

Studies that follow from what we present here should also have a more significant sample, compensating for one of the limitations of this investigation. We would like to point out that there were no significant differences between the groups in the oppositions that we tested, except between the conditions of “digital reading” versus
“sequential reading”, be it in the “natives” or in the “immigrants” group.

Conclusions

The results of this investigation demonstrated the absence of differences between the groups, particularly at a structural level, which, from our perspective, attributes extra responsibility to future information services and professionals, since as they deal with/manage information (as a product, a service, and a system), they prepare contexts that impact the users, namely in the development of intelligence, i.e., in their capacity to receive information, deal with it and produce efficient answers (Marina 1995).

Although libraries are no longer the premier place of access to information, they will be (in fact, they already are) places that condition and predispose subjects to acquire knowledge, generating states of pleasure, realization, and fulfillment of an objective. This place is intended for a user with an increasing difficulty to memorize read information when subjected to simultaneous stimuli, and with bigger concentration difficulties, with an impact on interpretation. We believe they developed a “get in, get the answer, get out” informational behavior (Thompson 2013).

In this sense, to predict the definition of informational behavior, the manner in which it will develop knowledge, and how the future reader will use that knowledge becomes a challenge, but it foresees a change in how information professionals must be prepared, and how they must think about information products and services. With regard to their qualification, it is expected that they should be well-educated, well-informed, able to deal with both human and technological issues, able to understand the essence of their work, not just as “manipulation” of materials, but as “serving” people in an ever more complex environment in terms of “noise” as well as of capacity of choice. This “new” information professional is not just a provider of information, but also an educator.

Regarding their role in the creation of information products and services, we hope that these professionals should be able to ensure unrestricted access to geographical and temporal spaces, conceiving and testing new means of information organization and retrieval: meaning and/or search for the source of meaning. This requires that they change their attitudes to the way they conceive, develop and implement information products and services. The logic of the “pull” environment should be substituted by the principle of the “push” environment. In practical terms, this means that information products and services should focus on personalization. Push logic requires a vision, a definition and a conditioning of the informational object even before we begin the search for it.

References


Figure 1: Representation of the block design of fMRI

Figure 2: Results (Difference of Groups)

Statistics: p-values adjusted for search volume

<table>
<thead>
<tr>
<th>p</th>
<th>c</th>
<th>q(MTE,corr)</th>
<th>q(ER,corr)</th>
<th>k</th>
<th>P(unc)</th>
<th>q(MTE,corr)</th>
<th>q(ER,corr)</th>
<th>k</th>
<th>P(unc)</th>
<th>t</th>
<th>P</th>
<th>k(unc)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.000</td>
<td>6</td>
<td>0.750</td>
<td>1</td>
<td>0.750</td>
<td>0.000</td>
<td>0.993</td>
<td>0.955</td>
<td>35</td>
<td>0.001</td>
<td>3.21</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>1.000</td>
<td>0</td>
<td>0.750</td>
<td>1</td>
<td>0.750</td>
<td>0.000</td>
<td>0.993</td>
<td>0.955</td>
<td>35</td>
<td>0.001</td>
<td>3.20</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>1.000</td>
<td>0</td>
<td>0.750</td>
<td>1</td>
<td>0.750</td>
<td>0.000</td>
<td>0.993</td>
<td>0.955</td>
<td>35</td>
<td>0.001</td>
<td>3.20</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>1.000</td>
<td>0</td>
<td>0.750</td>
<td>1</td>
<td>0.750</td>
<td>0.000</td>
<td>0.993</td>
<td>0.955</td>
<td>35</td>
<td>0.001</td>
<td>3.20</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>1.000</td>
<td>0</td>
<td>0.750</td>
<td>1</td>
<td>0.750</td>
<td>0.000</td>
<td>0.993</td>
<td>0.955</td>
<td>35</td>
<td>0.001</td>
<td>3.20</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>1.000</td>
<td>0</td>
<td>0.750</td>
<td>1</td>
<td>0.750</td>
<td>0.000</td>
<td>0.993</td>
<td>0.955</td>
<td>35</td>
<td>0.001</td>
<td>3.20</td>
<td>0.001</td>
<td>0.001</td>
</tr>
</tbody>
</table>
Figure 3: Reading condition TpB–TpA (1)
Figure 4: Reading condition TpB–TpA (2)
Abstract
This paper offers the results of an examination of library catalog analysis, a bibliometric approach in evaluating scientific research in Social Sciences and Humanities based on quantitative analysis of the presence of monographs in online library catalogues. The test was carried out in 2015-2016 in a large number of Italian and foreign libraries, highlighting the nonhomogeneous distribution of monographs in libraries. The survey demonstrated that calculating the presence of books in library holdings could not be considered a reliable indicator in research assessment. Moreover, the paper examines in depth recent tools for monograph search (discovery) and the Semantic Web standards for author identification.

Introduction
Research assessment is a scientific task that entails significant social aspects, as it aims at evaluating the scientific activity of each researcher, involving the activity of academic departments and the cultural policies of academic authorities.

The evaluation of research in Social Sciences and Humanities (SSH) is based on the analysis of books as the product of research, currently adopting the strategy of expert peer review. Henk Moed (2005) reported the problems that arise in SSH fields using bibliometric counts based on citations. Instead of computing the citations received by articles that scholars publish in journals, D. Torres-Salinas and H. F. Moed (2009), A. J. M. Linmans (2010), and H. D. White et al. (2009), autonomously proposed library catalog analysis1 as a bibliometric tool for research evaluation in SSH. Recently, Alesia Zuccala and Raf Guns (2013) again proposed the Libcitation count, previously suggested by White. Library catalog analysis is a new bibliometric approach that considers the number of copies of books in libraries using online catalogues, instead of counting the citations received by authors. It aims to evaluate the authors’ productivity in SSH and their scientific level.

The new bibliometric indicator for research evaluation in SSH proposed through library catalog analysis is a matter of importance in the knowledge organization (KO) field, as it involves library acquisition procedures and implementation of OPACs.

Objectives
The aforementioned in-depth analysis carried out covered a limited range of scientific fields: Economics (Torres-Salinas and Moed 2009), History and Literature (Zuccala and Guns 2013). Beyond these, the literature does not offer a wide empirical

1 According to Linmans, the correct definition is “Library Holdings Analysis”.

Maria Teresa Biagetti, Antonella Iacono, Antonella Trombone

Testing library catalog analysis as a bibliometric indicator for research evaluation in Social Sciences and Humanities
analysis of the potential and limitations of that strategy over a broad range of disciplines. Therefore, a deeper examination of scientific fields other than those tested by the authors mentioned is required.

This paper presents the concise results of an Italian survey on the diffusion in Italian and foreign library holdings of books published by Italian scholars in two scientific fields: history of books, bibliography, library science on one hand, and history of political institutions on the other. The survey, sponsored by the Italian Agency for the Evaluation of the University System and Research (ANVUR), was carried out in 2015 and 2016 with the purpose of verifying the possibility of using the results of analysis as an indicator in SSH research assessment.

Among the objectives of the research, was an analysis of the nature and operation of discovery tools in order to understand if these tools can affect the qualitative and quantitative outcomes of research into monographs. Discovery systems are software that appeared on the market around 2009 with features similar to the federated search tools created for electronic resources. Moreover, the survey aimed to highlight the possibilities offered by the new standards in the Semantic Web framework for the identification of authorships, which is a crucial point in research evaluation.

**Methods**

Authors who proposed the library catalog analysis mostly used unit catalogues such as WorldCat (Torres-Salinas and Moed 2009; Linmans 2010). In our survey, we preferred instead to verify the presence of books in a set of OPACs we chose as representative of different environments: 9 foreign libraries and 13 Italian libraries, including national libraries and libraries of prestigious European and US universities. Therefore, our work was not restricted to offering only quantitative results, but rather acquired a qualitative dimension. We set up a database of monographs published between 2000 and 2014 selecting them from the Italian SBN collective catalogue. The database includes 563 units: 279 in history of books, bibliography and library science, and 284 in history of political institutions.

Concerning the evaluation of discovery systems as tools for the research of monographs, a sample of monographs from the database used for our survey was also selected for testing discovery systems. The survey analyzed only a sample of the two scientific fields’ monographs: history of books, bibliography, library science and history of political institutions; then the results were compared with those of the same titles carried out in the respective libraries’ online catalogues.

**Main results**

Our survey produced several important results. First, the analysis of the holdings of Italian and foreign libraries allowed us to have a wide view and to compare different realities and make an articulate analysis. Our examination revealed a very mixed picture
and a substantial difference in the presence of monographs in the libraries in which the surveys were completed. Second, we stressed that key elements conditioning the inclusion of monographs in library holdings are numerous, vary according to the type of library and are often influenced by contingencies. Third, in evaluating the inclusion of books, we considered the incidence of gifts. None of the authors here mentioned, who proposed the library holdings as a reliable indicator of the scientific level of monographs, considered the incidence of gifts in library holdings.

Table 1 presents an aggregate view of the distribution of monograph publications in Italian and foreign libraries. Table 2 gives the detailed distribution of Italian monographs in national and academic foreign libraries, while Table 3 shows the detailed distribution of the same in national and academic Italian libraries. It is to be noted that Italy has 9 national libraries. In Table 3 includes details of 7 national libraries, as the central national libraries in Florence and Rome were not included, as they are mandatory repositories according to the Italian legal deposit act. Regarding the presence of monographs by Italian scholars in national and academic foreign libraries, the Harvard University libraries revealed the highest level of presence.

Table 1: Presence of monographs in the Italian and foreign libraries

Table 2: Presence of Italian monographs in national and academic foreign libraries

Table 3: Presence in national and academic Italian libraries

<table>
<thead>
<tr>
<th>CS</th>
<th>PZ</th>
<th>TO</th>
<th>NA</th>
<th>BA</th>
<th>VE</th>
<th>MI</th>
<th>CAG</th>
<th>VEA</th>
<th>RMS</th>
<th>UTO</th>
<th>UMC</th>
<th>USM</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>13</td>
<td>87</td>
<td>30</td>
<td>93</td>
<td>100</td>
<td>127</td>
<td>62</td>
<td>68</td>
<td>306</td>
<td>320</td>
<td>327</td>
<td>170</td>
</tr>
</tbody>
</table>

CS: Cosenza National library; PZ: Potenza National library; TO: Torino National and University library; NA: Napoli National library; BA: Bari National library; VE: Venezia National library; MI: Milano National library; CAG: Cagliari University library system; VEA: Venezia University library system; RMS: Roma Sapienza University library system; UTO: Torino University library system; UMC: Macerata University library system; USM: Milano State University library system.

In our survey, we found that occurrences are not always the result of conscious choices, but are influenced by several factors, such as cooperative policies for acquisition of publications, purchasing plans, approval plans and patron driven acquisition or particular forms of giving. Approval plans are a widespread acquisition method in libraries, based on the development of collections’ profiles and formal agreements signed between libraries and service providers. Title choices and purchase decisions are also often guided by users through a service called patron driven acquisition by means of which orders are generally suggested using next generation online catalogs.

The distribution of monographs in libraries analyzed in our survey is not homogeneous, whether in Italian or in foreign libraries. With respect to the field of history of books, bibliography, library science, one of the most significant results of our test was evidence that monographs on the history of books are preeminent in Italian and foreign libraries.

Another important result arising from our survey was its study of the operation of discovery systems as tools for research into monographs. A discovery system is a
centralized index of contents and metadata resulting from multiple different sources, which is combined with a multi-functional research system (Christensen 2013). It is presented as the solution to the problem of separately querying multiple indexes of collections that libraries give access to. The latest discovery tools perform a federated search (Wanga, Mi 2012) among all bibliographic library collections, even though their main feature is a central index which contains bibliographic records metadata, digital content in full text, metadata of institutional collections and online resources which the library gives access to through an indirect subscription, since they are included in the discovery system licence.

Web scale discovery systems thus allow users to query the various resources that libraries make available to them through a single access point. Discovery tools are index-based systems: all database contents, both local and remote, are re-indexed by discovery, including those subscribed to by the library with commercial agreements and catalogue data. Indexing and retrieval systems in the research phase are not clear, and there are no standards regulating this process, which remains, together with relevance criteria, completely outside librarians’ control (Breeding 2015).

Analysis of discovery tools use show that they are consulted above all as a catalogue, so as to look for publications held by a library; they are also widely used to download journal articles already known and frequently searched using Google Scholar (Moore 2016). However, the accuracy characteristics may not be the same as those obtainable by querying the search interfaces of the individual databases connected to the discovery. The discovery system offers less precision even than online catalogues and searching for a known document, that is to say a publication whose bibliographic identity is known can be problematic. In the same way, researchers may prefer to query native database indexes directly because they need precise answers, which do not require exploratory research (Ellero 2013).

A sample of monographs of the database used for our survey was also selected for testing discovery systems. Accordingly, we carried out a comparison between the outcomes of using the discovery system and the results of queries from libraries’ online catalogues.

The issues taken into consideration show a percentage of misalignment between the data from the catalogue and the discovery system. It is not possible to define the causes of this mismatch between data, which could depend on the software adopted and the settings defined for the attribution of relevance criteria to publications, for example.

The results showed that searches through online catalogues are to be preferred for quantitative and qualitative assessments, in particular for the comprehensibility of the document retrieval techniques, entirely managed in the field of library and information science.

Moreover, our survey highlighted the currently inaccurate management of identifiers
of entities registered in bibliographic records. Authority control is a crucial point for purposes of scientific evaluation. It enables precise identification of each entity of interest (author and publication) ensuring the correct attribution of works to their respective authors. The traditional authority files provide standardized access points for the authors of monographs, but often overlook the authors of articles in scientific journals, partly for economic choices, partly due to factors such as the preference given to new discovery tools for research (Martin 2015). This makes the task of philologically reconstructing the scientific production of an author very complex, because of the difficulty in obtaining results from a catalogue that associates monographs with contributions in periodical publications or with individual contributions in monographs. This condition could change with the implementation of the new bibliographic standards and protocols designed specifically for the Semantic Web: RDA and BIBFRAME. Resource Description and Access (RDA 2013) application requires a systematic process of ‘entity identification’, since every bibliographic data is univocally identified by the URIs and linked to other documents in the Web of data. Identifying entities with dereferenceable elements is thus the premise for creating a network of relationships between entities. The BIBFRAME model\(^2\) aims to replace MARC with a new extensible format, flexible and integrated with the Web and part of the Web of data. This huge change has important implications for controlled vocabularies (Southwick et al. 2015). Disambiguation and identification of authors are relevant issues in the Semantic Web where a different model is adopted from our bibliographic entity-relation model (E-R), such as the Resource Description Framework (RDF) model of triples and graphs. RDF application for structuring and modelling bibliographic information radically changes the identification process, which will be able to represent entities that appear in a variety of names in various resources (Peponakis 2016).

The collaborative nature of the Semantic Web introduces a new distributed vision of authority control, which no longer depends on national agencies, but involves subjects of various different types.

**Conclusions**

In our survey we analysed the presence of monographs in a range of unit catalogues and OPACs belonging to national and university libraries in Italy and abroad. Thanks to a deeper analysis of different library types, it has been possible to offer a broader picture and to better verify the possibility of using that indicator. We demonstrated also the effect of monographs presented as gifts in library holdings, verifying the number of monographs received as donations in particular by the Italian academic libraries of our sample. Moreover, our analysis was extended to include an accurate examination of the

\(^2\) https://www.loc.gov/bibframe/.
issues that arise using discovery systems in place of online catalogues, and the problems connected to the unambiguous identification of entities, which could be resolved with the implementation of Semantic Web standards. The latter analysis constitutes an important part of the testing procedure with regard to library catalog analysis, as we updated the study carried out by Torres Salinas – Moed, White and Linmans, and broadened our examination to include the modern tools (discovery and Semantic Web technologies) that in 2008-2009 were not yet available.

A comparison between the two scientific fields, history of books, bibliography, library science on one hand, and history of political institutions on the other, highlights that the latter shows a less widespread distribution compared to the former. This difference also persists in Italian libraries.

The results of our survey highlighted that the criterion of the presence of monographs in libraries cannot be adopted as a reliable indicator to establish the scientific value of monographs and, therefore, of their authors. However, it may be reasonable to claim that a check on the distribution of monographs in prestigious libraries, selected on the basis of well-balanced criteria, could constitute a means of reinforcing a positive or negative evaluation formulated following a well-managed peer review.

References


Linmans, Adrianus J. M. (2010). Why with bibliometrics the Humanities does not need to be the weakest link: indicators for research evaluation based on citations, library holdings, and productivity measures. Scientometrics, 83(2): 337-354.


Moed, Henk F. (2005). Citation analysis in research evaluation. Dordrecht: Springer.
Moore, Monica (2016). But is my resource included?: How to manage, develop, and think about the content in your discovery tool. The Serials Librarian, 70(1-4): 149-157.


Alexandre Fortier, Elaine Ménard

What do museum website users expect from linked open data?

Abstract
The use of linked open data promises to change fundamentally the ways in which users interact with information. For museums, finding museums how to encounter their audience in meaningful ways is key to fulfill their mission and optimize their investment. This study, which is part of a larger project examining the fundamental elements for the description of museum objects and their modelization using linked open data, aims to identify and understand the needs and expectations of museum website users in terms of descriptive data of museum objects and how linked open data would change their interactions with information offered on museum websites. Data gathered with a purposeful sample of 60 museum website users indicate that traditional physical metadata are uninteresting for the participants, who would prefer richer contextual information.

1. Introduction
Cultural and heritage institutions such as museums have a responsibility to make their “collections and all relevant information available as freely as possible” (International Council of Museums 2004). That responsibility, however, concerns a general access to collections, and museums have no obligation to afford a virtual access. For museums, developing and presenting virtual collections to the public constitutes considerable challenges which range from lack of financial resources to incompatibility of exchange between databases. The use of linked open data represents an interesting solution for cultural and heritage institutions to disseminate their information assets by establishing links between multiple data networks. Traditionally, search engines available on a museum website have the sole objective of performing a simple search on the museum’s own database. The search results are therefore limited to the content of that one database, which is often limited to basic descriptive elements such as the title of a work, the name of its creator or dimensions, and sometimes a picture of the museum object (Fortier and Ménard 2017).

The use of linked open data promises to change fundamentally the ways in which users interact with information. For museums, linked open data offer a feasible solution to overcome the obvious lack of compatibility between different databases. Linked data is not a defined standard per se, but a model for the publication of structured data on the Internet (Hooland and Verborgh 2014). The model consists of machine-interpretable triples in which a resource (the subject) has a specific relationship (the predicate) with another resource (the object). Elements in a triple use uniform resource locators for names, allowing a browser to locate a resource, no matter where it is stored. Open data, in turn, can be freely used, reused and redistributed. Linked open data effectively takes away the constraints of existing web approaches in which web users...
are forced to follow a pre-defined path chosen by the hosting organization to access the descriptions of museum objects. Thus, search engines can explore a collection of web resources to provide sophisticated and complete digital content. The desire to transmit and share digital content requires museums to envision a collaborative work logic, both amongst themselves and with other data provider.

The idea of using the web infrastructure for making metadata accessible in an interoperable way is not new, and projects have started to digitize cultural heritage materials through the use of semantic web technologies (Clough et al. 2008; Dekkers et al. 2009). Heath and Bizer (2011) provided a detailed explanation of linked data and a general state-of-the-art survey on existing data sources and solutions. Some link open data projects which specifically focused on museum objects have recently been launched (Oard et al. 2014). For example, Europeana, a digital library created in 2008 by the European Commission, brings together various digital resources (such as books, audiovisual material, photographs and archives) of national libraries from 27 countries. This organization published its own model of linked open data (Europeana Labs 2016). In 2014, the Smithsonian American Art Museum started publishing records of museum objects from their collections as linked open data, and consequently linking a growing body of open data published by organizations worldwide (Szekely et al. 2013). However, these projects are not yet widespread, and museums still hesitate to go in that direction.

For museums and other cultural institutions, despite the numerous challenges, having an online presence has become a requirement. Being digital, however, cannot stop at having a website, and museums must reflect on the ways in which they can develop strategies for becoming part of today’s digital society (Michaels 2018). Providers of digital museum resources often appear to be more focused on the mechanics of digitizing artefacts than on how people will use those digital resources once they are available. Simply piling more content onto a museum website is seldom the optimal strategy (Collection Trust 2015). To be successful at achieving their core mission of making their collections available to all using digital channels, museums must find how to encounter their audience in meaningful ways (Michaels 2017). Expectations of museum website users, however, are relatively unknown.

2. Objectives

This research project aims to identify and understand the needs and expectations of museum website users in terms of descriptive data of museum objects and how linked open data would change their interactions with information offered on museum websites. Specifically, this paper presents the needs and expectations of a group museum website users in terms of descriptive data of associated to museum objects of any kind as well as the ways in which linked open data can enrich their experience.
This piece of research is a portion of a larger research project that proposes to examine the fundamental elements for the description of museum objects and model them by using linked open data. More specifically, three objectives have been established for the overall project:

1. To understand what are the needs and expectations of museum site users in terms of descriptive data of associated to museum objects of any kind;
2. To define a model for the description of museum objects using linked open data that would be simple for museums to implement;
3. To strengthen data exchange networks amongst various cultural and heritage institutions.

3. Methodology

Over the last decade or so, several different methodologies have been proposed and tested for involving individuals in the interface and system design process. Amongst the numerous methodologies identified and described by Beheshti, Large and Clement (2008), this study proposes to use *Informant Design*, where the input of participants is required at various stages of the design process (Scaife and Rogers 1999). Participants are viewed as informants who can help researchers fill their knowledge gaps. *Informant Design* has been chosen for this study because it allows for gathering contributions from real users at crucial stages of model development.

In order to gain valuable information on the needs and expectations of museum site users, a sample of 60 museum website users was sought following a purposeful sampling strategy (Patton, 2015). Participants were asked to complete a questionnaire in which close- and open-ended questions were used to gather information on their needs and expectations regarding museum objects on museum websites, as well as how the possibilities brought by linked open data could change their interactions with information offered on museum websites. The data collection was completed by 15 individual follow-up interviews in order to have a deeper discussion about their needs and expectations. The interviews were audio-recorded and then transcribed. Qualitative data from the questionnaires and interviews were subjected to a content analysis based on a grounded theory approach (Corbin and Strauss 1990; 2015; Glaser and Strauss 1967) to discover emerging patterns in needs and expectations of museum website users.

4. Results

This sample is composed of 70% of women, aged between 18 and 55, and mostly in their thirties. All participants had university education, which is not representative of the general population, yet could make the results pertinent to museums where education is highly correlated with museum attendance (Smithsonian Institution 2007). The vast majority (92%) reported visiting a museum during the past 12 months, mostly
a few times during the year, and 88% reported visiting a museum website during the past 12 months. They mostly go to museums to do self-guided tours and visit gift shops. For 90% of the participants, the museum website is the first source of information to learn about a museum. Their main reason to visit a museum website is to seek entertainment, learn something, prepare a visit and learn about what is happening at a specific museum. On a museum’s website, they look for general information about the museum, such as schedule, location and prices, information about the museum’s collections and images of objects in the museum’s collections. A minority of them look for educational resources and virtual exhibitions. On a museum website, they expect general information about the museum, descriptions of objects in the museum’s collections and images of objects in the museum’s collections.

The majority (87%) of our participants reported using a museum’s website to look at objects in the museum’s collections. When they do, they look, in declining order, at images of objects, cultural context, dates associated, information about the creator of the object, history of usage and physical description. In terms of satisfaction related to the representation of collections on a museum’s website, only 30% of participants report being satisfied; 58% report being satisfied but would desire that museums provide more information about the objects in their collections; 8% report being unsatisfied. Very few of them (18%) are looking to find similar objects from a given creator or similar objects using another characteristic in the museum’s collection, and only 13% is looking to find similar objects using a given characteristic in another museum’s collection.

Qualitative results, however, indicate that participants find that museums, even major ones, generally present little information of interest about objects in their collection. Traditional physical metadata such as dimensions and material (or medium) are also seen as uninteresting. Information about cultural and geographical origins and period of objects appears to be of greater interest. Users would also be interested to find richer information about creators of objects and contextual information on use. Knowledge of linked open data was low amongst participants. When questioned about the possibilities that linked open data could offer, however, users expressed interest in discovering about similar objects in the museum’s collection or belonging to other museums.

Users were asked to rank the following statements in order of importance. A museum website must:

- Provide general information about its hours, events, services and facilities.
- Give access to information about the object in its collections.
- Give access to information about complementary collections available in other museums.
- Support the museum’s educational mission
Support the museum’s research mission

Very interestingly, the statement that is most often ranked in first position is that museums must give access to information about complementary collections available in other museums, which is a possibility that linked open data could support. In second and third places, participants placed the support to the museum educational and research missions. In fourth place was the general information, followed by the object in the collections, which are, according to responses to other questions, what they look for a museum website.

The majority of participants (75%) either agree or strongly agree that a museum is the sum of the physical museum and its website, although they are quite neutral with the idea that a museum’s website is as important as the physical museum itself. However, the majority of participants (80%) either agree or strongly agree that museums should put the emphasis on developing more web content, which should be relevant to a large proportion of the population (98% of participants either agree or strongly agree). Finally, 91% of participants either agree or strongly agree that museums should work strategically to develop their digital communication and dissemination of their collections online.

5. Limitations

It is essential that some limitations should be kept in mind when looking at these results. In this study, respondents were restricted to people who (1) were already visiting museum websites, and (2) chose to answer a survey about museums and museum websites. The purpose of user studies is not that, necessarily, any one slice of data is generalizable to the whole, but rather that each slice of data accumulated through multiple studies over time helps illustrate an overall trend, in this case with in museum website visitation patterns. When viewed from that perspective, this study helps to illuminate the mindset of a small, but nonetheless important, slice of a population of museum visitors. This phase is a starting point to conduct more in-depth interviews that will allow to dig deeper into the needs and expectations of users of museum websites, and therefore develop a model that will meet these needs and expectations.

6. Conclusion

Museum website visitors are interested in looking at virtual collections. They want images of objects, which have a significant presence on many museum websites, but they also want complementary information about cultural context, creators and history of usage, which are mostly non-existent (Fortier and Ménard 2017). Participants in this study would like museums to provide more information about their collections. However, they rank last that a museum must provide information about the object in its collections. For them, it seems that a museum should give priority to information that supports its education and research missions, such as rich narrative and complementary
information. This goes along the lines that a museum and its website are complementary, and that the website should not simply provide what is already accessible in the physical museum. Most museums appear to provide on their website the exact same information about the physical description that is available next to the object in the museum itself (Fortier and Ménard 2017). Last, participants strongly believe that museums should give access to information about complementary collections available in other museums, yet only 13% report looking to find similar objects using a given characteristic in another museum’s collection. One simple explanation would be that very few museums currently offer that possibility.

Museum websites should inspire visitors to experience the museum in person. In turn, people who visit a physical museum should feel inspired to visit the museum’s website, using the website as a bridge to learn more about the museum and its collections. It seems that, so far, museums have mastered the first part: websites, which are generally rich in general information about the museum, collections and specific exhibitions, are good at inspiring website users to visit the museum in person. If visitors in physical museums feel inspired to visit the museum’s website, however, they are often left with no more than what they saw on site.

Creating links between different databases offers a range of possibilities that will affect the ways in which users interact with information. The use of linked open data offers unprecedented context for enriching museum objects descriptions with existing metadata records with links to semantically related resources. Object descriptions could be augmented by adding to the data provided at a museum various elements from other databases. This is done independently of the language used to describe museum objects since links can be created between databases supporting multiple languages. With the integration of text and multimedia (3D images, sound explanations, etc.) this will constitute a benefit for certain groups of users with specific information needs (e.g., visually impaired could benefit from audio descriptions). The voice of users, however, remains relatively unheard in many initiatives using linked open data, and this project proposes to fill that gap, to ultimately provide a model that will meet their needs and expectations.

References


José A. Moreiro-González, Virginia Ortíz-Repiso

What is happening about KOS in Spain: scientific production analysis, 2000-2017

Abstract
This paper presents an analysis of scientific production from Spanish institutions in respect of knowledge organization systems and semantic vocabularies. The data sources are papers and communications indexed in the Web of Science (WoS) and Scopus during the years 2000-2017. The methodology has initially sought to identify, in the case of the works published in WoS, the subjects and temporary limitations of retrieval. The subjects were limited to the terms KOS and each of the types of vocabularies recognized in ISO 25964-1. 2011. The search was limited, according to documentary type, to articles and communications, as well as according to the geographical space belonging to the Spanish institutions. The search was further refined by requiring the presence of the search terms in the title, keywords or abstracts of the analyzed references. In the case of Scopus, the search categories are different and a relevant retrieval was found to be more difficult. It forced us to operate in a more indirect way. The results are shown in tables, and our comments reflect the institutional, personal and thematic characteristics of selected journals and conference papers, first comparing their international context, then comparing their internal profiles. The search was refined in WoS by the category Information Science and Library Science, by type of document, by country and by date. The same search terms were used in Scopus, and were filtered by Arts and Humanities (ARTS) and by Social Sciences (SOCI). The results are presented according to the different vocabularies and KOS used in the selected records. The most notable trend is marked by the opening up of the KOS to engage in the information and business management of very dynamic lines of activity, such as online reputation, energy efficiency, e-commerce, business accounting, finance and banking, organic agriculture, and industry.

Introduction
The analysis of Spanish productivity in respect of KOS by using two databases, WoS and Scopus, provides a necessary guarantee of relevance of the documents extracted. The journals and conference proceedings that appear in these two indexes guarantee a scientific communication that is of greater visibility and also, therefore, widely applicable, with the result that the contents can have a greater impact. All journals and contents have passed peer reviews (Bar-Ilan, Levene, Lin 2007). Due to the international circulation of their articles, WoS and Scopus are the best means of access to the paths followed by the different specialties, and, certainly, the best means of testing the state of research efforts in Spain in relation to knowledge organization. (Moreiro 2017). The fact that our research is limited to the articles and communications indexed in WoS and Scopus means that the results obtained cannot be considered sufficiently complete. For this reason, it would be advisable to give attention to books and PhD dissertations, and even take into account other types of publication, which would ensure a more complete vision. But to do this would be a longer task, and would not ensure the exactitude of consultation provided by WoS and Scopus. We have therefore confined our study to the scientific journals and conference proceedings
included in WoS and Scopus between the years 2000 - 2017. Our aim in this study is to:

- Select and identify the contribution of Spanish researchers to KOS, from a theoretical point of view and also in practical terms.
- Analyze and evaluate the data obtained in order to test the degree to which the information is interdisciplinary, and also analyze authorship and institutional origin.

**Methodology**

The research was confined to the main collection of the database Web of Science (WoS), by combining the terms that are defined as types of KOS in the international standard ISO 25.964-1 (2011) and making use of the logic operator OR. The terms were truncated, in order to include the extensions of the terms in the search. When necessary, a search by exact phrase was carried out, in order to avoid noise during recovery. The following search terms were used in the field TS (abstract, title and keywords): KOS; “knowledge organization system*”; ontologie*; “information representation*”; thesaur*; “controlled vocabular*”; “semantic network*”; folksonom*; “indexing language*”; “subject representation*”; taxonom*. The results of the search were refined by selecting the category Information Science and Library Science, according to types of document (articles or proceedings papers) and according to country (Spain), and selecting period of time 2000-2017. The search was limited to the indices: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI.

In Scopus the same combination of search terms was used and the same search procedure was followed, searching in title, abstract and keywords for articles or proceedings papers. Since it is not possible to search by the sub-category topics Information Science and Library Science, it was necessary to filter the search by arts and humanities (ARTS) and by social sciences (SOCI). Subsequently, in order to assign the sub-area Library and Information Science (LIS) each title was assigned the topic classification of Scimago Journal Ranking (SJR). In order to achieve this, we created a special relational database, cross-referencing the registers obtained with the SJR.

Some of the registers recovered relate to topics that are tangential to KOS, but we have accepted them, since they are included in the print-out of recovered data. We have only excluded 7 records: in these cases, we could not obtain evidence of Spanish authorship, or evidence that the publications had been prepared in Spanish institutions. Another 5 documents were also excluded, because the topics were not related to KOS. In the final analysis, the total number of works selected was 229. The accuracy of data recovery was high (0.02) after setting a limit in Scopus for LIS, and after excluding 5 out of the total of 229 documents selected between both databases.
Results and conclusions

In order to present the results obtained, tables and figures have been prepared to make visualization simpler. The source of all data presented is always the WoS and/or Scopus. The first three tables reflect the general results obtained by the search, before proceeding to refine the registers obtained for the country Spain. Lastly, the references are presented in the international standard format ISO 690-2013.

In WoS, the final result was that 174 documents were recovered, although following data selection this figure was reduced to 171. In Scopus, the final result was that 279 documents were recovered; however, only 143 relate to Library and Information Science or are relevant to our research, and these documents are the ones that have been made use of for the purposes of our study. The registers that coincided in both databases were compared, and it was found that 85 documents were included in both platforms. This represents a coincidence of 49.70% in WoS and a coincidence of 59.44% in Scopus.

<table>
<thead>
<tr>
<th>Source</th>
<th>N. Articles</th>
<th>N. Proceed.</th>
<th>TOTAL</th>
<th>Selected</th>
<th>Coincidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>WoS</td>
<td>136</td>
<td>38</td>
<td>174</td>
<td>171</td>
<td>80</td>
</tr>
<tr>
<td>Scopus</td>
<td>229</td>
<td>50</td>
<td>279</td>
<td>143</td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>229</td>
<td>50</td>
<td>279</td>
<td>143</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>229</td>
</tr>
</tbody>
</table>

It was found that the United States generates a quarter of the total of 2,612 documents that represent the total of world production recovered in the search, after application of the refinement by Information & Library Science. However, this total is lower than the combined contributions of the European countries (Spain, England, Italy, Germany and France), listed in the results obtained after application of this first refinement.

The Spanish contribution in this sector is particularly noteworthy, if we bear in mind the size and populations of the countries classified in the following table. Spain occupies second place among the European countries listed, and fourth place worldwide. We calculate that Spain is the source of 7.01% of all documents recovered.

<table>
<thead>
<tr>
<th>Country</th>
<th>Records</th>
<th>% of 2612</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>678</td>
<td>25.96</td>
</tr>
<tr>
<td>England</td>
<td>183</td>
<td>7.01</td>
</tr>
<tr>
<td>China</td>
<td>178</td>
<td>6.82</td>
</tr>
<tr>
<td>Spain</td>
<td>174</td>
<td>6.66</td>
</tr>
<tr>
<td>Italy</td>
<td>141</td>
<td>5.40</td>
</tr>
<tr>
<td>Brazil</td>
<td>135</td>
<td>5.17</td>
</tr>
<tr>
<td>Canada</td>
<td>125</td>
<td>4.79</td>
</tr>
<tr>
<td>Germany</td>
<td>113</td>
<td>4.33</td>
</tr>
</tbody>
</table>
The contribution of each of the countries listed among the top twelve countries classified is very significant. Each of these countries is the source of at least 4% of world production. Particularly outstanding is the presence of an Asian country, China; also Canada, in North America; while in South America Brazil stands out; and in Oceania, Australia.

Table 3: Number of documents per institution (Source: WoS 2017)

<table>
<thead>
<tr>
<th>University</th>
<th>Records</th>
<th>% of 2612</th>
</tr>
</thead>
<tbody>
<tr>
<td>Univ Wisconsin</td>
<td>36</td>
<td>1.38</td>
</tr>
<tr>
<td>Univ Illinois</td>
<td>30</td>
<td>1.15</td>
</tr>
<tr>
<td>Megill Univ</td>
<td>29</td>
<td>1.11</td>
</tr>
<tr>
<td>Univ Washington</td>
<td>25</td>
<td>0.96</td>
</tr>
<tr>
<td>Wuhan Univ</td>
<td>25</td>
<td>0.96</td>
</tr>
<tr>
<td>Univ Fed Minas</td>
<td>24</td>
<td>0.92</td>
</tr>
<tr>
<td>Nanyang Technol</td>
<td>23</td>
<td>0.88</td>
</tr>
<tr>
<td>Univ Alcalá de</td>
<td>23</td>
<td>0.88</td>
</tr>
<tr>
<td>Univ N Carolina</td>
<td>22</td>
<td>0.84</td>
</tr>
<tr>
<td>Kent State Univ</td>
<td>21</td>
<td>0.80</td>
</tr>
<tr>
<td>Univ Granada</td>
<td>21</td>
<td>0.80</td>
</tr>
<tr>
<td>Columbia Univ</td>
<td>20</td>
<td>0.77</td>
</tr>
<tr>
<td>Univ British</td>
<td>18</td>
<td>0.69</td>
</tr>
<tr>
<td>Univ Carlos III</td>
<td>18</td>
<td>0.69</td>
</tr>
<tr>
<td>Russian Acad Sci</td>
<td>17</td>
<td>0.65</td>
</tr>
<tr>
<td>Ionian Univ</td>
<td>16</td>
<td>0.61</td>
</tr>
<tr>
<td>Univ Fed Santa</td>
<td>16</td>
<td>0.61</td>
</tr>
<tr>
<td>Univ Manchester</td>
<td>16</td>
<td>0.61</td>
</tr>
<tr>
<td>Univ Pittsburgh</td>
<td>16</td>
<td>0.61</td>
</tr>
</tbody>
</table>

According to number of documents produced, the top 20 institutions worldwide are the following: in the United States, the universities of Wisconsin, Illinois, Washington, North Carolina, Columbia, Indiana, Kent State and Pittsburgh; in Spain, the universities of Granada, Alcalá de Henares and Carlos III, both in Madrid; and in Brazil, the federal universities of Minas Gerais and Santa Catarina. In addition, the Canadian universities of British Columbia and McGill, Wuhan University in China, the Ionian University on the Greek island of Corfu, and Manchester University in England, all stand out. The fact that Spain has three universities among the top 20 is an accurate reflection of the interest that exists in Spain for this line of investigation.
Spanish production of documents shows some unevenness, although it appears as a continuing progression upwards during the period 2000-2017. The sudden boom observable in 2006 could be due to the publication of the first international regulatory standards relating to semantic vocabularies (Z3919 2005; BSI Group 2005). In Spain, the most productive years can be seen to be 2008, 2013 and 2016, showing 9.19%, 10.34% and 13.79% of documents respectively. These boom years may be compared with other less productive ones, where considerable downturns in production can be observed, although these were quickly readjusted. In 2010 and 2011, for example, production reached only 6.3% and 5.17% respectively: this could be considered as an immediate reaction to the cut in the funding allocated to research, due to the economic crisis. The 24 documents corresponding to 2016 allow us to conclude that this line of research continues to attract a healthy interest in Spain. If we compare the situation in Spain with the evolution of publication worldwide, it can be seen (Figure 2) that publication peaked in 2008, although from that year onwards a downturn in production commenced, only starting to rise again (though not to recover previous levels) in 2016.
In Spain, the institutions that have been shown to be most prolific in terms of publication during the period examined are the following: the University of Granada, the Carlos III University in Madrid, and the Universities of Alcalá, Zaragoza, and Murcia, the Consejo Superior de Investigaciones Científicas [Higher Council for Scientific Research], the Polytechnic University of Valencia, and the University of Alicante. This shows that there is a clear correlation between the ranking of the most prolific authors in Table 4 and the ranking of the universities where they work.

In respect of the journals in which the greatest number of papers are published, the Spanish journal *El Profesional de la Información* is the most prominent, with 21.40% of total production (229). The journal *Knowledge Organization* ranks second (9.17%), followed by *International Journal of Metadata, Semantics and Ontologies* (this journal is only listed in Scopus) with 10.91%; *Online Information Review* (8.73%), *Information Processing Management* (7.42%); and, finally, *Information Systems Management* and *Revista Española de Documentación Científica*, both representing 6.98% of production.

In view of the considerable differences in the classifications in WoS and Scopus and with the aim of reflecting the content of the documents as clearly as possible, we have made our own classification, and the following analysis:
Table 5: Interdisciplinarity (Source: the authors)

<table>
<thead>
<tr>
<th>Interdisciplinarity</th>
<th>WoS Ocur.</th>
<th>Scopus</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Library and Information Science</td>
<td>171</td>
<td>143</td>
<td>314</td>
</tr>
<tr>
<td>Library and Information Science (Applications and Theoretical contributions)</td>
<td>68</td>
<td>39</td>
<td>107</td>
</tr>
<tr>
<td>Computer Science</td>
<td>53</td>
<td>4</td>
<td>57</td>
</tr>
<tr>
<td>Health Sciences</td>
<td>16</td>
<td>4</td>
<td>22</td>
</tr>
<tr>
<td>Social Communication Science</td>
<td>15</td>
<td>10</td>
<td>25</td>
</tr>
<tr>
<td>Education</td>
<td>13</td>
<td>9</td>
<td>22</td>
</tr>
<tr>
<td>Geography</td>
<td>10</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>Linguistics</td>
<td>8</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Agriculture</td>
<td>7</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>Engineering</td>
<td>6</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>Business Organization</td>
<td>6</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>History</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Politics</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Research Activity</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>The Humanities</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Environmental Sciences</td>
<td>4</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Law</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Tourism</td>
<td>3</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Music</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Psychology</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Fine Arts</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Mathematics</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Physics</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Sociology</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Politics</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The areas included in the above list serve to illustrate the interdisciplinary nature of the documents selected. There is nothing strange in this, considering the epistemological roots of knowledge organization that are an inheritance of the transdisciplinary nature of information and its applications. (Holland 2008). This effect is certainly increased during the period selected for our study, since it coincides with the increase of digitization in systems and processes, the spreading of communication via social network applications, and the development of computer networks.

Due to the particular characteristics of the searches carried out, all the documents are included in the category LIS. Again, the KOS papers that investigate applications, and the theoretical studies also included in LIS, take first place in the ranking, with 39.7% in WoS and 27.27% in Scopus. But these documents are in fact very closely related to Computer Science studies when they investigate ontologies or conceptual thesauri design and management, similarity measurements between concepts, automatic classification, semantic structure of data, schemes, etc. These are all hybrid
areas of research, and so the methods used, the areas of interest and ways of working are very closely related in both.

We should bear in mind that KOS articles will be included in LIS, even if they are published in journals and congresses relating to other thematic categories – just as we have seen can happen the other way around - because LIS has an instrumental character (SCIMAGO 2005). This occurs particularly in the documents listed in Scopus, because it is impossible to carry out a manual recount, due to the considerable dispersion, in Scopus, of LIS contents relating to other areas and categories.

In the table, it can be seen that the 229 documents selected show a distribution that is dispersed over 24 areas of knowledge. And despite applying extremely broad classification criteria such as academic areas and productive activity, we are able to identify a considerable mix of knowledge in the documents, as already noted by Rumrill, Fitzgerald and Merchant (2010). This provides us with sufficient authorization for describing KOS research as, first and foremost, an intersecting field.

Comparing the results of the columns relating to WoS occurrences and Scopus occurrences, the difference in the incidence of Computer Science is particularly striking: 53 to 4. The explanation is to be found in the way each of these systems classifies areas and categories: an autonomous existence is allowed to LIS in WoS, but not in Scopus. No doubt this required a more demanding process of analysis during filtering, in Scopus, resulting in the exclusion of documents of mixed content. The greater number of occurrences in WoS also leads us to deduce that seven of the categories included in WoS do not qualify for entry in Scopus even though five of the categories only appear four times, and five of them only appear once. KOS is used in Education in a very particular way, in order to classify learning objectives.

Ideally, of course, all the references included in each of the documents recovered ought to be listed. In addition, an analysis should be included, even if brief, of the content of each document. However, the limits of our research only enable us to enumerate the various different KOS taken into account in the selected documents.

<table>
<thead>
<tr>
<th>KOS types</th>
<th>WoS</th>
<th>Scopus</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ontologies</td>
<td>52</td>
<td>28</td>
<td>80</td>
</tr>
<tr>
<td>Thesauri</td>
<td>39</td>
<td>31</td>
<td>70</td>
</tr>
<tr>
<td>Taxonomies</td>
<td>37</td>
<td>5</td>
<td>42</td>
</tr>
<tr>
<td>Vocabularies</td>
<td>25</td>
<td>10</td>
<td>35</td>
</tr>
<tr>
<td>Theoretical</td>
<td>11</td>
<td>7</td>
<td>18</td>
</tr>
<tr>
<td>Abstracting</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Folksonomies</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

Comparing WoS with Scopus, almost twice as much attention has been allotted by
WoS to Ontologies. And in WoS far more documents relating to Taxonomies and Classification Schemes are listed, while they barely feature at all in Scopus. We may perhaps give the same explanation for this as we gave earlier: the presence in both databases of Computer Science, because of its interdisciplinary nature. In other types of KOS, however, we found a more equal balance between the two databases. The fact that Thesauri features so strongly in both databases is a clear indication of its prestige and recognition in LIS. The listing of three documents in Abstracting is due to the fact that the content of these documents combines with Ontologies or Vocabularies.

Ontologies, conceptual Thesauri and linked Taxonomies are the three types of KOS that appear most frequently in the documents recovered in our research. We can deduce from this that both LIS and Computer Science give special importance to these categories. However, articles are still being published in the traditional way about thesauri and classifications. Theoretical studies are also allowed considerable importance in both databases, especially around the time that international standards were published, in 2005. 49 documents combine two types of KOS in their texts; and there are two instances where three types of KOS are combined in the same document. Thus we see an obvious fact of KOS reflected here: both Taxonomies and Thesauri have to function ontologically. Furthermore, it has been demonstrated that the various different types of KOS are not radically distinct from each other, since, if the ambiguous references in vocabulary and relations models are resolved, all the types are affected. (Tudhope; Binding, 2006). Knowledge organization in network makes this proximity possible, using vocabulary coding schemes – NKOS - together with SKOS data models.

References


Posters
Challenges in management and Knowledge Organization of documental heritage: key factors in the methodology of diligent search for orphan works

1. Introduction and objectives

The UK Intellectual Property Office (2015) defines orphan works as “copyright works where one or more of the right holders are unknown or cannot be located”. In the same line, the European Directive 2012/28/EU (2012) indicates that “a work or a phonogram shall be considered an orphan work if none of the rightholders in that work or phonogram is identified or, even if one or more of them is identified; none is located despite a diligent search for the rightholders having been carried out”. In this context it must be clarified that “before a work or phonogram can be considered an orphan work, a diligent search for the rightholders in the work or phonogram, including rightholders in works and other protected subject-matter that are embedded or incorporated in the work or phonogram, should be carried out in good faith” (European Directive 2012/28/EU).

The following specific objectives are indicated for the project:
- To present the reference context of orphan works as elements making up documental heritage.
- To define and delimit the concept of diligent search processes for orphan works from the perspective of the “Library and Information Science” domain.
- To identify and define the elements and key terms related to the management and organization of the knowledge involved in a diligent search process.
- To analyse the core components serving as the basis for the systematization and generation of record files for the said diligent searches for the declaration of works in this category.

2. Methods

The key elements defining the methodology on which our paper is based are as follows:
- **Scope**: analysis of the diligent search process for orphan works and the delimitation and study of its components for normalization and systematization.
- **Research techniques used**: content analysis for data sources and specialized documentation; benchmarking or examination of the best institutional practices related to the subject of our research; normalization techniques for procedures and the process management approach for organizing knowledge applied to the
management of orphan works.

— **Data sources:** key documentation and legislative provisions in this matter, technical standards and reports, specialist databases and web sites of the leading institutions in this area, scientific and technical reports generated by research teams currently working directly on this subject and other supplementary data sources such as specialist articles in the matter under study.

3. Main results and conclusions

The study starts by delimiting the concept of orphan works as a set of document types (printed works, such as books or periodicals, cinematographic or audio-visual works and phonograms, as well as works incorporated into another work or in a phonogram, such as images or photographs, for example) forming part of the documental heritage preserved in libraries, archives, film or audio institutions or other bodies for the conservation of cinematographic or audio-visual heritage, public-service broadcasting organisations, among others. The fact that these are copyrighted works whose rightsholders have not been located, poses a series of challenges and difficulties about how to approach the search for information about the work as well as the identification and location of the rightsholders of the works in question.

Starting from this conceptual delimitation, this study presents a list of member states in which we have identified the existence of best practices related to diligent search procedures which have to be carried out before a work can be declared orphan in the countries of the European Union where Directive 2012/28/EU has been implemented. In line with this analysis, a methodological and conceptual proposal is set out for a normalized diligent search process so that this can be applied and adapted by any documental heritage institution proposing to embark on a process of this type. Regarding this methodological proposal, we present an identification and definition of the elements and terms related to the management and organization of the knowledge involved in a diligent search process. These include, on the one hand, those related to the identification, description and systematization of the data sources and reference documentation, as well as, on the other hand, those related to the existence of databases for the registration of orphan works.

Finally, with respect to what could be defined as a process of documentation for this diligent search, we present a proposal for the identification of the basic informative/descriptive elements that must be reflected on what could be called the search record of the diligent search performed as evidence of the process carried out.

In short, the goal of this proposal focuses on responding to the aim of any documental heritage institution being able to count on practical guidelines to be able to cope, on the one hand, with the challenge of conducting a diligent search and, on the other, with the challenge of recording its results. It should be recalled that, in accordance with current
legislation, these records must be completed and kept by the said cultural institutions so that this information can be validated by the competent national authority in each member state as the prior step before declaring the work as an orphan on the database created and managed for this purpose by the European Union Intellectual Property Office (EUIPO).

As a preliminary conclusion, it is possible to advance that we have detected the existence of a relationship between the *de facto* declaration of orphan works on the EUIPO database and the degree of development of instruments, procedures and soft-law tools available to documental heritage institutions. In this line, we feel that our paper can be viewed as a contribution for the support of orphan work diligent search methodologies and their registration. In consequence, the management and organization of knowledge around the normalization of the diligent search process for orphan works is viewed as a key success factor that can help to resolve the challenges posed to documental heritage institutions relating to this research subject.

References

A Connotative trust-based paradigm to minify societal challenge(s) in Knowledge Organization Systems

1. Introduction

Knowledge creation, knowledge sharing and idea generation are all critical to enhancing productivity and adaptation to global changes in this age of information. Personal information is the currency of the 21st century digital economy. Additionally important is the fact that information has no transaction costs and does not decrease in value when the supply increases. Contrary to the laws of economics, it may even increase in value with greater supply. This in our opinion suggests the need for better ways to share information such that when internalized - absorbed and understood by the human mind - it becomes knowledge (Liew 2007). One barrier to knowledge sharing is the huge societal divide in terms of the myriad of culture and beliefs in the world. This trust divide spells a challenge for society, which can hinder access to knowledge on a global scale (Hanson and Kararach 2011).

Interestingly, trust comes from mutuality which attracts reciprocity. This makes trust an issue to consider since people are very likely going to show aversion when they interact in the global society. This cannot be ruled out in the current world, which has become even more interconnected, and both the receiving and sharing of knowledge continue to be from (so many) different sources (Ali et al. 2015). We believe that potential problems arising from the foregoing can be minified. This paper proposes attributes as parameters that can strengthen existing trust models towards better online interaction with satisfactory minimization of trust (Hendrikx et al. 2015). Therefore, the potential of a trust-based paradigm is used to propose a trust collaborative mechanism that ensures no break in mutuality. This is because when “non-breaking” mutuality is achieved, trust as a societal challenge will be controlled and minified.

2. Literature

The issue of trust, particularly in a social interactive situation, requires trust-based collaboration. Some researchers have tried to make contributions in this regard (Clarke et al. 2010; Truong et al. 2011). For Clarke et al. (2010), every participant should communicate directly only with his own trusted peers. While Clarke et al. (2010) developed the Darknet model for effective trusted peer communication, Truong et al. (2011) presented a push-pull-clone model for trust-based collaborative editing with contract deployed over F2F network. In their model, contracts were specified when data are shared between friends, yet how participants in the network will cope with untrusted
peers was not provided for. The provision made in this current work uses the concept of social trust. This concept highlights the fact that social trust entails a situation whereby change in trustworthiness is inevitable.

However, the metric of trust was used to calculate the trustworthiness of other peers. Tan *et al.* (2011) also developed a distributed trust model based on social network principles. In addition, Xiong and Liu (2004) developed the PeerTrust model, which is useful for quantifying and comparing the trustworthiness of peers based on a transaction-based feedback system. Each of the foregoing models used either reputation or direct experiences (personal trust) to determine trustworthiness. The trust-aware model developed in this work handles trustworthiness from the perspective of personal trust, reputational trust and recommendation trust.

3. Methodology

This paper will take a cue from the theoretics of trust-aware collaborative modelling based on a Friend-to-Friend Network (F2FN) scenario. As a result, a trust-aware system that uses an F2FN collaborative scenario will be presented. This will be highlighted as an effective management platform to test trust with respect to a robust collaborative knowledge sharing paradigm. The system will use the trustworthiness of collaborators explicitly to be able to grant access. We believe that trust is mutable. Thus, how the trustworthiness of peers is upheld as a priority (Aderibigbe *et al.* 2016) will be demonstrated by how the system updates itself during and after collaborations. Aside from the ability of the system to update itself, this updating will be followed by trust computation, whose result can be looked up and updated accordingly in preparation for fresh or continuous collaboration. Metrics like reputational accuracy, convergence and effectiveness will be used to assess the level of trust among collaborating and knowledge sharing peers. This will also be used to validate the specific tasks that are needed to ensure trust during knowledge sharing.

We envisage the possibility of churn, which is normal in an F2FN situation. We also foresee a situation in a social media network context, when communication will be direct with peers, and most times without considering the level of trust they have towards one another. It will therefore be imperative to make trust data available to aid trust computation. The distributed hash table in synergy with the chord technique and the symmetric replication technique will thus be introduced. Using an architectural approach, the trust-based system will be highlighted to show the function of peer update of trust after a collaborative session. Based on the foregoing architectural provision a prototype system will be presented to demonstrate the presence of a significant improvement in reputational accuracy and convergence when deployed to manage trust using an F2F scenario.
4. Expected results

This paper will show the possibility of minifying the societal challenge of lack of social trust in a knowledge sharing context. The plausibility of the foregoing possibility is in the fact that the trust-aware model’s prototype system dealt with trust issues. We submit to the fact that this can inspire a connotative inference for better knowledge organization systems (KOS). This implies that where trust issues are rife (e.g. in social media and other related networks), trust can be minified. The overall corollary from the presentation so far will be the trust-aware system as a prototype. Its application will be used to unequivocally infer the imperatives that are needful in minifying trust issues in KOS.

5. Conclusion

It will therefore be concluded that trust as an obvious societal challenge can be minified using appropriate techniques, cognizance of an appropriate scenario and KOS context(s).

References


Hanson, K. T., and Kararach, G. (2011). The Challenges of knowledge harvesting and the promotion of sustainable development for the achievement of the MDGs in Africa.


The Contribution of research on information quality to Knowledge Organization

Introduction and objectives

Since the 1990s several proposals have been published about frameworks to be used in the context of information quality in information science. According to Eppler and Wittig (2000), an information quality framework must provide four things: first, a systematic and concise set of criteria according to which information can be evaluated; second, a scheme to analyze and solve problems related to information quality; third, a basis for measuring the information quality and enable proactive management; and finally Fourth, a conceptual map for the research community that can be used to structure a variety of approaches, theories and phenomena related to information quality.

In order to contribute to future research in the area of information quality, this study will present a literature review about the frameworks proposed in the last 16 years. This review intends to complement previous studies and contribute to the body of research specialized in the quality of information for the generation of knowledge, for the development of new studies in the area, as new frameworks that could be applied in specific contexts.

Frameworks for information quality

To establish an overview about information quality and its frameworks, in April 2017 we conducted basic searches using the terms “framework” and “information quality” in the following databases: BRAPCI (Base de dados de Periódicos em Ciência da Informação, a Brazilian information science database), Web of Science (WoS) and Library & Information Science Abstracts (LISA). The choice of LISA and BRAPCI, was because they have a direct relation with the field of information science and our purpose is to realize the different frameworks proposed in this area of knowledge. The search using Web of Science was filtered for just the category "Information Science & Library Science".

Table 1 presents the frameworks found in the literature from 2001 to 2016. It indicates articles in which framework proposals are presented, including the author name, year of publication, title and a summary of each article.
<table>
<thead>
<tr>
<th>Title, authors and year</th>
<th>Sumary</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIMQ: a methodology for information quality assessment (Lee et al., 2002)</td>
<td>Consists of a generic model for evaluation of information quality that contains an application model, a questionnaire and techniques for the analysis of the results.</td>
</tr>
<tr>
<td>Evolutinal data quality: a theory specific view (Liu and Chi, 2002)</td>
<td>Generic model that consists of four steps related to data quality that generate information: collection, organization, presentation and application.</td>
</tr>
<tr>
<td>Measuring information quality in the web context: A survey of state-of-the-art instruments and an application methodology (Eppler and Muenzenmayer, 2002)</td>
<td>Framework that allows the assessment of the multiple dimensions of information quality in the context of the Internet or Intranet.</td>
</tr>
<tr>
<td>Getting better information quality by assessing and improving information quality management (Caballero; Gómez and Piattini, 2004)</td>
<td>Framework composed of two elements: a quality management model structured at the levels of maturity and a methodology for evaluation and improvement of information quality.</td>
</tr>
<tr>
<td>A semiotic information quality framework: development and comparative analysis (Price and Shanks, 2005)</td>
<td>Generic approach to evaluating the quality of information based on semiotics, providing a framework containing quality categories and their criteria integrating objective and subjective quality visions.</td>
</tr>
<tr>
<td>Operations quality of data and information: teleological operations research-bases approach, call for discussion (Gackowski, 2005)</td>
<td>Generic theoretical model, where all attributes are determined and evaluated by the purposes and circumstances of the operations, as well as the quality aspects of the people involved.</td>
</tr>
<tr>
<td>A framework for information quality assessment (Stvilia et al., 2007)</td>
<td>Generic framework consisting of typologies of comprehensive problems, related activities and a taxonomy of the dimensions of information quality organized on a systematic basis in theory and practice.</td>
</tr>
<tr>
<td>The impact of diverse information systems environments on information quality - a design science approach (Foley; Helfert; and Elwood , 2010)</td>
<td>Generic framework that can be adapted to different environments with a specific focus on information systems.</td>
</tr>
<tr>
<td>Information quality on the World Wide Web: development of a framework (Kandari et al., 2011)</td>
<td>Framework that aims to measure the quality of information on the Web from the perspective of the user.</td>
</tr>
<tr>
<td>Judging the quality and credibility of information in internet discussion forums (Savolainen, 2011)</td>
<td>Framework for assessing the quality of information by specifying criteria that are used to judge quality and credibility in Internet discussion forums.</td>
</tr>
<tr>
<td>A methodology to evaluate important dimensions of information quality in systems (Todoran et al., 2015)</td>
<td>Framework divided into three stages: first, the system is decomposed into its elementary modules, second, each module is analyzed individually and third, the previous two steps are used to measure the quality of the system as a whole.</td>
</tr>
<tr>
<td>A model for the evaluation of data quality in health unit websites (Leite; Gonçalves and Teixeira, 2016)</td>
<td>Evaluation model for the quality of information in institutional sites of health units.</td>
</tr>
</tbody>
</table>
Conclusion and future work
Managing the quality of information is critical within any context since information is linked directly to knowledge generation. Increasingly, information users demand quality information and thus call for new solutions that meet their needs.

The objective of the survey of frameworks for information quality presented in this poster was to offer the researchers and scholars of information science and knowledge organization a synthesis of the proposals made in the last sixteen years and present a set of references that can be consulted in order to propose new models for evaluating information quality.

As a possibility for future research, it is suggested that the survey on frameworks for information quality, be carried out in databases other than solely in the area of information science, since this area has its own guidelines for dealing with information quality.

References


Creation of a domain ontology in CIDOC CRM OWL format using heterogeneous textual data related to industrial heritage

An important social issue in the cultural heritage domain is related to collection, analysis, publication and enhancement of collective history and memory of stakeholders, whether spoken or written. Formalization of cultural heritage information presents a real challenge because of the data diversity and incompleteness. Moreover, such data are heterogeneous and can be found in different sources, online or offline: databases, libraries, museums, press papers, stakeholders’ expertise, etc. In addition, we are witnessing a prodigious rise in the volume of digital and physical contents describing this heritage and the increase of the production power associated with dissemination techniques, at different scales, especially at the regional level. This variety of resources brings many problems such as data documentation, representation, integration and interoperability within the same knowledge base. Most of the attempts to resolve semantic interoperability problems focus on standardization and development of shared structures like FRBR, FRBRoo, CIDOC CRM etc. Among such technologies, the CIDOC CRM is a Conceptual Reference Model specifically designed for modeling cultural heritage domains. This model provides a common meta-data schema making it understandable and interrelated by implicit and explicit relationships. Such ontological schemas are meant to promote a common data understanding about cultural heritage by providing a common and extensible semantic framework in which all information could be mapped.

In order to assist domain experts in producing and providing digital contents, our project, called DENIM, aims at providing a knowledge-based representation that interconnects all data, by using semantic web technologies. Its originality is a multidisciplinary approach to assist stakeholders, experts and non-experts for knowledge discovery specific to their heritage, through the extraction, structuring and visualization of knowledge extracted from heterogeneous digital corpora. According to UNESCO, which has contributed significantly to the definition of the heritage (UNESCO 1954, 1970, 1982), and then to the International Committee for the Conservation of Industrial Heritage (TICCIH 2003), industrial heritage can be defined as:

- Material assets: buildings, machinery, equipment, workshops, factories, processing and refining sites, shops, production centers and social activities related to the textile industry;
Immaterial assets: memories, events, festivals, collective images, intellectual production transmitted by know-how which can be a succession of gestures dictated and displayed in production centers.

In our work, efforts are concentrated on modeling the domain of stakeholders, the spatial and thematic entities, which belong to each of these assets. Thereby, we first present the main concepts and relations that we keep from the CIDOC CRM model to represent industrial heritage. Among the concepts which are the subject of an equivalence between the textile industry and the CIDOC CRM, we can mention materials (E57 Material), dimensions (E54 Dimension), states (E3 Condition State), current and past locations (P53 has former or current location/is former or current location of), inscriptions (E34 Inscription) and representations carried by the object, its date of creation, documentation and publications devoted to the subject or an associated web page as well as iconic reproductions, or the rights attached to documents (E30 Right). Information actors can also be specified, for example the name of the specialist who proposed the attribution to a document. Depending on the needs, the CIDOC CRM may be used partially or, instead, can be associated with new headings, such as the type of object that has been added thanks to combinations of classes and properties used to process a thesaurus (E55 Type). Geographical data also has an important place, represented in the CIDOC CRM with the class E53 Place, a concept which is associated with E47 Spatial Coordinates, to indicate its coordinates. The latter are used in particular for searching and displaying data on an interactive map, enabling the answers to be modulated chronologically and accompanied by statistics. The inclusion of CIDOC CRM by experts allows not only the simultaneous interrogation by Internet users of several large information sources but also facilitates the integration of new partners, who have only to establish the equivalence between their data and the selection of standardized notions coming from the textile domain. The use of these meta-data also paves the way for new exchanges with other initiatives that respect the principles of this standard.

We adopt a three-step methodology that allows semi-automatic construction of a semantic representation of the studied domain from heterogeneous documents (cf. Figure 1).

Firstly, we collect and formalize the history by interviews with stakeholders. In addition to the information collected, we also exploit a web mapping/visualization of stakeholders organized by their type (Kergosien et al. 2015).

Secondly, our methodology operates the identification and the extraction of information related to industrial cultural heritage from heterogeneous textual documents (interviews, numerical documents from libraries, newspapers, etc.). The proposed approach combines lexicon projection with text mining methods to improve
the identification of relevant information. Lexica of spatial entities initially cover regional municipalities of Hauts-de-France, France. The lexicon of the domain’s stakeholders was built semi-automatically with experts. It includes the full names and acronyms used for the naming of the actors - for example UIT (Union of Textile Industries) or IFTH (French Textile and Clothing Institute). To create a thematic lexicon, existing specialized resources defined by experts (Joconde created by French museums, Rameau created by the National Library of France, Wiktionary, etc.) were analyzed and filtered manually. From this step, the result is a set of 2,000 terms related to the textile domain. From this first lexicon, we have defined text mining approaches based on the Word2vec algorithm in order to identify new terms from the processed corpus. Specifically, the method transforms sparse vector representations of words (provided by the lexicon and the analyzed corpus) into a dense, continuous vector space, enabling the large-scale identification of similarities between words and phrases, based on their context.

For example, we reproduce below the set of terms associated to wool. In black are the terms already present in our first lexicon, in blue are the terms added to the lexicon such as suggested by the corpus, in orange are those which should be validated by the field expert, and in red are the erroneous associations.

- wool: cotton, drapery, carding, combed, lining, linen, blouse, silk, satin, jute, cheviotte, oozing, fabric, velvet, pure, woolen, skein, pasha, tarpaulin, rattling, crocheted, washing, crossed, thistle, plain, spun, ecru, cretonne, overcoat, waste, fluff, potash, satin, millinery, canvas, tow, garment, crossing, mix, tennis.

The main purpose of the method is to build a semantic representation of the studied domain as precisely as possible. The indexed documents are structured in XML MODS format1 (cf. Figure 2), defined by the Congress Library in the United States. This standard is a compromise between the complexity of the MARC format used by libraries and the extreme simplicity of the Dublin Core metadata.

Finally, we present a first version of the automatically built ontology, in the OWL format using the CIDOC-CRM model as conceptual basis, to merge together all our lexica. During this phase, it is important to filter the CIDOC CRM model to obtain a sub-model with the relevant concepts and properties.

The experiments were conducted on a corpus with thousands heterogeneous documents (newspaper articles from LaVoixDuNord, documents with meta-data from libraries, and interviews) related to the Textile Industrial Heritage (TIH) located in Hauts-de-France, France. The ontology built is tested and validated by experts using Protege (cf. Figure 3 and Figure 4).

Regarding future perspectives, we plan to extend this work by designing a generic

1 http://www.loc.gov/standards/mods/.
and semi-automatic approach for building semantic representation related to industrial heritage. In addition, we propose to test our method on heterogeneous data related to industrial and mining heritage, collected within the framework of the MemoMines project.

Figure 1: Organization of the information extraction method for creation of the ontological database
Figure 2: An indexed document in XML MODS format

```xml
<?xml version="1.0" encoding="UTF-8"?>
<mods>
  <titleInfo>
    <title>Quatre cents élèves sont venus visiter le mini-musée des Anciens de la Toissee</title>
  </titleInfo>
  <typeOfResource>text</typeOfResource>
  <genre>article de presse</genre>
  <originInfo>
    <dateIssued>24/12/2010</dateIssued>
  </originInfo>
  <subject>
    <country>France</country>
    <state>Nord</state>
    <city>Tourcoing</city>
  </subject>
  <subject>
    <country>France</country>
    <state>Nord</state>
    <city>Roubaix</city>
  </subject>
  <subject>
    <country>France</country>
    <state>Nord</state>
    <city>Wattrelos</city>
  </subject>
  <subject>
    <topic>acrylique</topic>
  </subject>
  <subject>
    <topic>textile</topic>
  </subject>
  <subject>
    <topic>laine</topic>
  </subject>
</mods>
```
Figure 3: Ontology instantiation using the transformation module

Figure 4: Excerpt of the domain ontology based on four heterogeneous documents (MODS XML, PDF) using the Protege Software (Musen et al. 1995)

References
Juan Bernardo Montoya-Mogollón, Sonia María Troitiño

Diplomatic forensics: a necessary historical review for the analysis of the born-digital record

The document is understood as the predominant means of communicating human activity in all its variety. The transformation of the support has created new challenges for its conservation in the different stages of history: stone, clay, papyrus, parchment and paper, among others. These supports have elements in common: 1. they are intelligible directly by the human eye; 2. their conservation and preservation may be possible even in difficult conditions; and 3. when a modification or change of information is made intentionally, its detection is easily identifiable.

The contribution of diplomatic and forensic science has established from its beginnings, theoretical and practical elements to validate and prove the authenticity of records produced in different institutional environments. Diplomatic validation is done by analyzing documents’ internal and external nature, that is, in a critical and in-depth way, it analyzes the information that constitutes the document and its support, as well as additional information that can be fixed in the document. Forensic science, in turn, reinforces the methods of diplomatic validation, ensuring the documentary chain of custody so that it can be used as a source of proof in legal, administrative and other judicial scenarios.

The emergence of the two sciences mentioned refers essentially to the Middle Ages, a historical moment in which the states of Europe began to solidify. The official and solemn documents issued by religious chancelleries were sources of assurance to add the political power of the states in formation. Moreover, the falsification of diplomas (official documents), began to be a common practice to legitimize this political force. Therefore, it was necessary for the church to establish standards and methodologies to analyze the authenticity of these documents, first undertaken by an "empirical" diplomat.

Later, in the 17th century, the Benedictine Jean Mabillon developed a methodology to analyze the nature of documents that were believed to be false. In 1681, he published De re diplomatica libri sex, thereby establishing the scientific character of diplomatics and paleography. From that moment, different faculties of law in Europe expanded study into the authenticity of diverse documents, especially those of a legal nature. The importance of the forensic field was decisive in linking areas of knowledge such as diplomatic science and forensic science, although today only diplomatics is regarded as an autonomous science.

The objective of the present work is to carry out a historical review of diplomatic
and forensic science from its beginnings, to then analyze the resurgence of these two applied sciences in the context of the digital record. Digital diplomatic science assists in verifying the authenticity, reliability and completeness of the digital record, taking into account both its identity and its integrity. Forensic science, on the other hand, establishes rules to verify that the record is not corrupted at any stage of the documentary life cycle.

The methodology used was theoretical-descriptive and exploratory, taking into account the studies of both classical and modern authors. Studies related to areas such as information sciences and knowledge organization, archival science, forensic sciences, diplomatic science and law. It analyzes concepts little addressed in the field, such as digital forensic sciences linked to computer science to try to assist the process of verification and identification of the authenticity of digital archival digital documents.

The results of the research focus on providing the born-digital record with sufficient guarantee for it to be used as a source of administrative, archival, legal, and legal evidence, among others, by analyzing both the internal and external characters within digital systems that produce, conserve, preserve and give access to the user. Likewise, the way diplomatic science and forensic science can subsidize elements to ensure the authenticity of documents at any stage of the documentary life cycle.

The relevance of the present research is based on the fact that we currently need sciences that contribute to the conservation and preservation of the information and knowledge that is being produced in different contexts. It is being observed that digital documents have a high risk of becoming obsolete in the short term and therefore the intervention of sciences such as diplomatic and digital forensics is urgently needed to preserve the documentary collection as a means to keep alive the memory of society.

References


*Digital Records Forensics Project*. Available at: http://www.digitalrecordsforensics.org/.


The French military documentary system to anticipate health risk: content and information classification

Introduction
French armies must manage international crises in a constantly changing environment. They must also anticipate these crises and make quick decisions facing health events that may impact the forces in operation. It is to meet this objective that the Medical Intelligence service of the Center for Epidemiology and Public Health of the Armed Forces (CESPA) has been given the task of developing a medical intelligence system that collects information on health risks in military operation (Boutin 2004).

Objectives
This article aims to present the medical intelligence system developed to respond to the different needs of sanitary information of the forces and military decision-makers. It is a system fed by fresh health data via an automatic or semi-automatic process, whereby information is synthesized and organized in such a way that it is categorized and easily searchable.

Methodology
The system developed is powered by a documentary monitoring process in 6 steps with feedbacks: information content definition, automated collection on databases (e.g. MedLine, ToxLine, Web of Science, etc.), human and automated analysis of the information, classification and distribution in the system. The selection of documentary content is validated with users (physicians, health policy makers, etc.). The documents are classified according to the classification of the National Library of Medicine (NLM) and named according to a naming procedure (Tanti 2012). The selected and classified resources are capitalized in a documentary system which is based on four different information systems (IS) described in this article.

Results
The documentary system developed integrates access to multiple sources of information such as databases and hypertext. It is available on a secure Intranet. It is mainly based on four IS defined by Boulogne as “an organized unit, inter-connected procedures, methodologies, elements of organization, human resources, software and computer equipment allowing to collect, store, structure, insert, extract, move, control, post, exchange (transmit or receive) and to communicate the selected information in form of text, images, sound or coded data for the use of exploitation, and management.
of an organization” (Boulogne 2004).

**Presentation**

The first IS is called BEDOUIN. It provides the military health service directorate, medical advisers of military commands as well as physicians, pharmacists, and veterinary surgeons at unit level with updated information on potential health risks for deployed French forces. The documents in the IS are indexed by keywords, in natural language and by using the Medical Subject Headings MeSH thesaurus. The consultation is done either by hypertext navigation starting from an index by country, (animal and human) diseases or toxicological risks or via a search engine. BEDOUIN provides 189 files relating to different countries. It also contains 893 scientific documents on 233 human diseases and 162 veterinary diseases and 1978 toxicological files.

The second IS is called REDUVES, which contains Research and Development information. It is made up differently in terms of appearance and its content is complementary to that of BEDOUIN. It informs the Military Health Service directorate and medical advisers of military commands about the latest scientific publications on agents of medical importance, in particular biological agents and agents that can be used for the production of weapons of mass destructions, and corresponding research activities. It provides R&D files giving an overview on preventive measures, diagnostic procedures, therapeutic assets and reports on documentary awareness. As in BEDOUIN the consultation is done either by hypertext navigation starting from an index by risk and agent, or via a search engine. REDUVES currently contains 46 files on agents posing health risks, 26 R&D files, 26 reports on documentary awareness, 20 works and reviews as well as 2,955 scientific source documents.
BOUGAINVILLE aggregates rough scientific documents on agents of biological, natural and provoked risks, chemical substances, whether of industrial or military nature, and allows the permanent development of BEDOUIN and REDUVES. In terms of organizing and gathering information by topic, documents are classified by generic and specific terms, in folders and subfolders (classification by risks encountered outside the territory – biological, chemical and industrial – and by agents – smallpox, anthrax, etc.). Documents are indexed by keywords, in natural language and by using the MeSH thesaurus, by first author and by year. To enable the rapid location of documents, groups of documents or passages of text which answer specific questions, an electronic document management tool was installed. It allows the user to query the whole base, via simple or advanced search, by specific fields (title, summary, etc.), by free or controlled key words, and via boolean operators. BOUGAINVILLE, permanently evolving, contains 30,900 documents including 25,800 on biological risks, 2,070 on chemical risks, 280 on radiological risks, 610 on veterinary medicine related risks and 2,150 documents relating to different countries.

The last IS is IntraCespa, which deals with reports, notes, bulletins of epidemiological surveillance of the French forces and other internal documents relating
to epidemiology and public health issues of the French military. INTRACESPA documents are consultable by author, topic or type, in multi-criteria mode, simple or advanced search but also by arborescent navigation.

Figure 4: INTRACESPA homepage

Example of use

In the next Ebola epidemic, French military authorities and physicians at unit level regularly consulted the documentary system to receive real time information on the global evolution of the medical crisis. They also consulted scientific documents on the virus, research publications on its mechanisms of pathogenicity, as well as the daily progress in vaccine and treatment developments. Physicians were able to seek information and ministerial recommendations, in particular within the framework of epidemiological surveillance and aspects of protection campaigns of military and civilian populations (Chatelet 2015).

Conclusion

The system developed by the French military to monitor health risks in operation is regularly updated by a documentary monitoring process. The documents of the system are searchable for users, especially for preparing deployments on theaters of operations or to respond to health crises, such as during the recent outbreak of Ebola and Zika. In the near future, it will be consultable by satellite, under extreme conditions. It has proven its daily use and utility in crisis situations. It would now be interesting to qualitatively and quantitatively evaluate its effectiveness and efficiency.

References


Gamification as a system for developing knowledge in the classroom: a proposal based on an educational innovation project

1. Introduction and objectives

Gamification is a term that emerged in the business context in 2008, and that consists in the inclusion of game elements in contexts different from those traditionally related to leisure and entertainment (Deterding, Dixon, Khaled and Nacke 2011). For Burke (2014) “gamification is the use of game mechanics and experience design to engage and motivate people to achieve their goals”. For Kim (2015) gamification goes beyond game design, since it means using its power to motivate people, attract their attention, involve them in an activity and even influence their behavior.

Gamification as a strategy can respond to different orientations that could be synthesized as follows: a) a recreational type, as employed in libraries in order to attract users, or b) a learning type, to motivate and transmit knowledge in the classroom.

The following provides a summary of a classroom innovation project proposing the development of a pilot gamification experience in subjects related to the management of information and documentation units.

The following specific objectives are indicated for the project:

- Application of gameplay scenarios in the context of the classroom for the creation of a game-like experience.
- Application of gamification techniques to develop knowledge and improve students’ practical learning, motivation and academic performance, with respect to their skillsets for management and planning of information and documentation units.
- Design of gamification proposals related to challenges in the real/professional world and to students’ core interests about the above-mentioned subjects.
- Fostering teamwork, participation, interaction and mutual enrichment of the various participants in the experience.

2. Methods and results

The project responds to the goal of developing an educational innovation experience offering support and assistance for structuring strategies based on gaming as an educational tool and as a system for the development of knowledge in the classroom. The aim is to increase motivation among the intended recipients of the project. It is a pilot experience supported by active teaching-learning methodologies and strategies applied to subjects related to the management and planning of information and
documentation units, but it could be exportable to other subjects/areas and/or to other additional scenarios.

In terms of methodology, the planning of an innovation project in the classroom has been based on the formulation of the following phases:

An initial phase known as “Proposal and selection”, which consists of an identification and selection of the gaming strategies and dynamics related to the managing and planning skillsets for information and documentation units.

A second phase designated “Planning, design and development of gaming strategies and dynamics” which covers the design of the gaming strategies and dynamics selected in the previous phase, their scheduling and their implementation and application in the classroom.

Thirdly, the “Documentation of gaming dynamics phase”, which involves the recording and registration of the gaming strategies and dynamics, the generation of a directory of standard records for each one, the publication of the documentation generated and, finally, the analysis and storage of the documentation resulting from the whole design and implementation of the gaming strategies and dynamics.

Lastly, a final phase of “Analysis and dissemination of the project’s outcomes”, including the analysis of the pilot experience and the lessons learnt, as well as the drafting of a final report and the presentation of the results at specialized events.

In terms of results, content has been selected and designed for the following activities:

– Teaching content: “The principles of scientific management”. Activities: Management Crossword (game based on a crossword with questions related to the subject), Hidden people (game based on clues to visually discover a reference author in relation to the content studied) and Kahoot! Trivial (online game of questions and answers about content faced by different teams. This game is based in the Karhoot! software, available in: https://kahoot.com).

– Teaching content: “Project Teams training”. Activity: The Nine Belbin Team Roles. The Belbin method is used to identify people's behavioral strengths and weaknesses through nine different clusters of behavior that are displayed in the workplace. In this adaptation, students identify the nine roles in their projects design.

– Teaching content: “Project ideas: competition”. Activity: Shark Tank. Shark Tank is an international reality television program which shows aspiring entrepreneur-contestants as they make business presentations to a panel of “shark” investors, who then choose whether to invest. In this adaptation the pupils show project presentations to “shark” teachers.

– Teaching content: “Staff Selection and Job Interviews”. Activity: Role-play. Students perform a role play with a script previously prepared by the teachers.
The session is recorded on video and is then studied in the classroom to see how it was developed.

Figure 1. Two examples of activities: crossword and hidden people

3. Conclusions
This educational innovation project model has been enthusiastically received by all participants, both students and teachers, who showed a high degree of motivation in using it. The activities planned and developed within the teaching contents have been well accepted with a high level of participation, which allows us to think about future actions that can even improve these gamification actions and, moreover, apply them to new contents. From the point of view of teaching, the activities have facilitated the practical understanding of the contents and a considerable improvement in the results and final evaluation of the students.

References


1. Objectives
This poster presents the implementation of a phenomenon-based knowledge organization system in the context of the ongoing design of a discovery system for our institutional research data repository. Besides the requirements for interdisciplinary information retrieval, there are challenges to overcome the limitations and shortcomings of existing metadata standards for research data management and library catalogues. To demonstrate how these issues can be handled, I will focus on the prospects and opportunities of knowledge organization for interdisciplinary research organization through ontology-based modeling and Linked Data technology. We want to achieve interoperability between institutional research data management, the institutional research information system and an institute’s library catalogue in order to integrate information about research objects, methods, data and related publications towards better information access to support the information needs of interdisciplinary researchers.

2. Approach
There are two steps to reach our objectives: first, integration of information; second, enrichment of the integrated information with facets for interdisciplinary knowledge organization.

2.1. Integrating a library catalogue, research information and research data
Bibliographic data is available for integration from the B3Kat library catalogue via SPARQL. The ISF (Integrated Semantic Framework) ontology of the VIVO research information system provides an opportunity to interlink research data and bibliographic data. Shortcomings of existing metadata standards, such as not using authority files or using them inconsequently, and inconsistent use of controlled vocabularies are solved in VIVO by enforcing role-based modeling of authorship, affiliation, etc. with authors, organizations, etc. identified via linked ORCID or GND records. We use CKAN as a datastore for our institutional research data repository because of its DCAT vocabulary, the RDF and Linked Data features and some useful view extensions. In addition to the Map View and Geospatial View extensions for georeferenced maps (from our GeoPortOst project) and spatial research data (like historical census data), we use CKAN’s data extension which supports the DOI registration service (also used in our
LaMBDa data portal).

2.2. Superimposing linked data with an interdisciplinary Knowledge Organization System

To build the faceted classification system, phenomena, methods, theories or paradigms are selected from the JEL classification system for the discipline of economics, subsets from GND subject headings converted into SKOS for the discipline of history and the European Thesaurus on International Relations and Area Studies for conflict studies.

The integrated records are superimposed with facets to classify research objects, methods used and theories as the subject of publications and research data from different disciplines. The simplest solution would be to use a subject classification property for each facet. As SKOS does not support facets (Isaac & Summers, 2009), i.e. the combination of concepts, n-ary relations (Noy and Rector 2006) would be needed for multiple facet classifications.

As an alternative to subject classification properties and n-ary relations, the NeDiMAH Methods Ontology (NeMO) (Benardou et al. 2010) was selected to model the digital research process more precisely. Comparable to phenomenon-based classification (Gnoli and Szostak 2007; Szostak 2008; Szostak et al. 2016), NeMO distinguishes between research goals and activities, research techniques, and research objects. Phenomenon-based knowledge organization systems “allow users to shift from one perspective or viewpoint to another, thus reflecting the multidimensional nature of complex thought. In particular, it should allow them to search independently for particular phenomena, for particular theories about phenomena (and about relations between phenomena), and for particular methods of investigation” (Gnoli and Szostak 2007).

3. Conclusion

The phenomenon-based faceted classification system builds the groundwork for the navigation and search system. The ontology-based information architecture meets the requirements of interdisciplinary researchers to find relevant information within the institutional research data repository. It also allows complex queries via SPARQL and integration with other repositories via Linked Data. The latter is planned in a large DFG-funded research infrastructure project in cooperation with the Bavarian State Library and other Leibniz research institutes as partners.

Finally, an outlook shows preliminary results of prototypical automatic faceted classification based on the KEA keyphrase extraction algorithm. KEA’s SKOS support (Medelyan and Witten 2006) is used to select matching entries from the faceted classification system. This controlled indexing approach tries to find information about research objects, methods, theories and research data in the abstracts of the respective...
research articles. The performance of KEA is compared to human indexers and first experimental results achieved with the GROBID (N)ERD tool for named entity disambiguation (Riondet and Foppiano, 2017).

References


This paper studies the influence of Open Access (OA) on institutional repositories. It summarizes a variety of Open Access models from both a scientific and general viewpoint.

The paper considers the causes for, and methods used to establish institutional repositories from KO aspects. In the introduction, mention is made of university policies relating to the publication in OA journals and deposits in an OA repository, particularly within their own institutions. The usability and possibilities of information retrieval are studied, together with problems relating to knowledge organization in these systems. The methodology used to explore how the organization of knowledge is expressed in these repositories, is also explained.

The papers aim to show some experiments to interlink the local and international institutional full text repositories – particularly the ones at universities, central databases and the IFLA Library repository.

The time has come for a paradigm shift related to scholarly publication and publishing. In this digital world, the community of research and academia can retrieve many things, but equally, much can also be lost. Libraries could respond to their obligations relating to the information and content services, where previously, this was not possible because the price of scholarly journals was very high (Bánhegyi 2003).

The paper presents and compares the solution of knowledge organization in three repositories:

1. Eötvös Loránd University's Digital Institutional Repository (EDIT)

EDIT stores the documents and archives of the oldest and biggest university in Hungary. Within the categories, users can search and browse collections. There are currently 32,590 documents in this repository. EDIT contains the following structure and search term categories:
2. HUNOR

There are a number of initiatives for creating institutional repositories in the country. The federation of Hungarian repositories is HUNOR – HUNGarian Open Repositories. The HUNOR consortium was established in 2008 by the 7 libraries of Hungarian higher education institutions and the Library of Hungarian Academy of Sciences to advance national Open Access practices. There are currently 24 member institutions. *Repositories of HUNOR consortium have different KO solutions.*

The members of HUNOR are dedicated to promoting Hungarian research both nationally and internationally and to achieve effective dissemination of scientific output through the implementation of a national infrastructure of Open Access repositories. HUNOR is coordinated by the National and University Library of the University of Debrecen (HUNOR leaflet 2012).

3. IFLA Library

The IFLA Library collates IFLA's (International Federation of Library Associations) digital resources in a convenient single online location for ease of accessibility, search and browsing. It was launched in summer 2013 with the IFLA World Library and Information Congress papers, and will continue to grow with the addition of existing and new resources. Currently containing 1,636 papers (http://library.ifla.org/), this repository uses the topical search term categories, based on sections and JITA system (http://eprints.rclis.org/cgi/search/advanced). JITA is a classification schema of Library and Information Science. It is used by E-LIS for indexing and searching. JITA is currently available in 38 languages including Hungarian. It is also accessible as Linked Open Data (http://aims.fao.org/vest-registry/vocabularies/jita-classification-system-library-and-information-science).
**IFLA Library uses the following schema and hits:**

JITA Classification System of Library and Information Science (63)

<table>
<thead>
<tr>
<th>Category</th>
<th>Subcategory</th>
<th>Hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.</td>
<td>Theoretical and general aspects of libraries and information (1)</td>
<td>GE. Staff (3)</td>
</tr>
<tr>
<td></td>
<td>Relationship of LIS with other fields (1)</td>
<td>GG. Curricula aspects (3)</td>
</tr>
<tr>
<td>B.</td>
<td>Information use and sociology of information (16)</td>
<td>GH. Education (6)</td>
</tr>
<tr>
<td></td>
<td>Use and impact of information (1)</td>
<td>GI. Training (4)</td>
</tr>
<tr>
<td></td>
<td>Information in society (5)</td>
<td>H. Information sources, supports, channels (14)</td>
</tr>
<tr>
<td></td>
<td>Information policy (1)</td>
<td>HB. Gray literature (1)</td>
</tr>
<tr>
<td></td>
<td>Information dissemination and diffusion (5)</td>
<td>HD. Rare books and manuscripts (2)</td>
</tr>
<tr>
<td></td>
<td>Information needs and information requirements analysis (2)</td>
<td>HE. Print materials (1)</td>
</tr>
<tr>
<td></td>
<td>Use studies (4)</td>
<td>HH. Audio-visual, Multimedia (2)</td>
</tr>
<tr>
<td></td>
<td>User studies (4)</td>
<td>HL. Databases and database Networking (1)</td>
</tr>
<tr>
<td></td>
<td>User training, promotion, activities, education (5)</td>
<td>HQ. Web pages (1)</td>
</tr>
<tr>
<td></td>
<td>CE. Literacy (2)</td>
<td>HR. Portals (2)</td>
</tr>
<tr>
<td></td>
<td>D. Libraries as physical collections (30)</td>
<td>HZ. None of these, but in this section (4)</td>
</tr>
<tr>
<td></td>
<td>DB. National libraries (2)</td>
<td>I. Information treatment for information services (14)</td>
</tr>
<tr>
<td></td>
<td>DC. Public libraries (2)</td>
<td>[IFLA Library homepage]</td>
</tr>
<tr>
<td></td>
<td>DD. Academic libraries (2)</td>
<td>IA. Cataloging, bibliographic control (2)</td>
</tr>
<tr>
<td></td>
<td>DF. Government libraries (6)</td>
<td>IB. Content analysis (A and I, class.) (1)</td>
</tr>
<tr>
<td></td>
<td>DH. Special libraries (4)</td>
<td>ID. Knowledge representation (2)</td>
</tr>
<tr>
<td></td>
<td>DI. Science libraries (2)</td>
<td>IE. Data and metadata structures (1)</td>
</tr>
<tr>
<td></td>
<td>DJ. Technical libraries (6)</td>
<td>IH. Image systems (5)</td>
</tr>
<tr>
<td></td>
<td>DK. Health libraries, Medical libraries (5)</td>
<td>IK. Design, development, implementation and maintenance (3)</td>
</tr>
<tr>
<td></td>
<td>DM. Museums (1)</td>
<td>IZ. None of these, but in this section (1)</td>
</tr>
<tr>
<td></td>
<td>DZ. None of these, but in this section (1)</td>
<td>J. Technical services in libraries, archives, museum (6)</td>
</tr>
<tr>
<td></td>
<td>E. Publishing and legal issues (1)</td>
<td>JF. Paper preservation (1)</td>
</tr>
<tr>
<td></td>
<td>EB. Printing, electronic publishing, broadcasting (1)</td>
<td>JG. Digitization (4)</td>
</tr>
<tr>
<td></td>
<td>F. Management (12)</td>
<td>JH. Digital preservation (3)</td>
</tr>
<tr>
<td></td>
<td>FA. Co-operation (8)</td>
<td>K. Housing technologies (2)</td>
</tr>
<tr>
<td></td>
<td>FD. Public relations (3)</td>
<td>KB. Library, archive and museum buildings (1)</td>
</tr>
<tr>
<td></td>
<td>FZ. None of these, but in this section (1)</td>
<td>KD. Vehicles (1)</td>
</tr>
<tr>
<td></td>
<td>G. Industry, profession and education (13)</td>
<td>L. Information technology and library technology (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LR. OPAC systems (1)</td>
</tr>
</tbody>
</table>
Main results

The analysis of repositories from a knowledge organization (KO) point of view highlights the consequences of a lack of, or unsatisfactory value of classification, and an incomplete information retrieval system. From this survey of repositories and systems, it can be seen that:

- there are not many subject-based systems in repositories
- there is some use of international systems, such as JITA, but there is no system similar to those in general use in Hungarian repositories
- KO systems are not cohesive, relating to the number of documents and the complexity of KO

There will be more repositories with increasing collections in the future e-society. Open Access will bring many positive changes concerning accessibility, but the principal conclusions consider the results of the search in repositories: the major problem being that the KO system has not yet been fully developed and standardized in the repositories.

References


1. Introduction

Interoperability is the exchange of information and its use without special effort on the part of systems (ALA 2000; Zeng and Chan 2004). This implies the possibility of different knowledge organization systems (KOSs) communicating with each other with no dependence on a precise actor except on an open standard such as a user interface (UI) (McCrees and Daniel 2017). This delineation is drawn on to present interoperability as a noesis or abstraction that is conceptualizable as a UI. KOSs source(s) and target system(s) have their own – independent – structures and characteristics, hence interoperability cannot be ignored. A UI can be used to hide all the details in each KOS structure and characteristics, and users would only have to interact with the UI since it integrates all the various KOS notwithstanding their structures, purpose and characteristics.

A KOS is layered; and consists of theoretics, structures, symbols, the digital aspect(s), and the application layer. The contribution of this work falls within the application layer (Gnoli 2013). This paper draws a connection between user experience (UE) and (current) interactivity attributes and reports on the possibilities their presence on a KOS UI means towards supporting interoperability for better information access. The aim is to combine the continuous disparate source(s) and heterogeneous target systems that Search Engines (SEs) rely upon - the example KOS highlights the need to find this link. For example, SE - particularly those of the Web – interact with key elements (Garud and Kumaraswamy 2005) (i.e. KO systems) such as thesauri, taxonomies, folksonomies and registers (Shiri 2014). These KO systems are incongruent and need to understand each other for the SE to provide satisfactory services to users. UE, Interactivity and other user attributes are provisions that are required to develop interoperable UIs (iUIs). For instance, a good SE should be able to collate users’ IN(s) as if it completely understands them. The UI can provide a standard for KOSs to connect with each other with no dependence on each other without the user knowing. This can only be possible if there exists a UI that promotes user attributes like UE and interactivity to the foregoing end. Thus, useful user experiences and robust interactivity can influence interoperability and facilitate access to information.

2. Literature review

In the KOS literature effort has been made to introduce interoperability within a single search apparatus (Zeng and Chan 2004). Existing UI were only able to cope with
the level of interoperability that was needed then, unlike now. The consensus in the
digital environment is for more possibilities in terms of information presentation and
access like Google that provides aggregate search platforms. However, it is necessary
to render the different KOS interoperable within a UI. If interoperable KOSs are to be
developed, it requires some level of cogitation, which is both technical and more
intellectually abstract (see Figure 1).

Figure 1: The Intellectual abstraction towards UI modelling

3. Method/methodology

In order not to violate the general principles of UI design especially “user control
and freedom” we are conscious of the fact that UE attributes may be unmanaged. With
the attribute of interactivity (at present), such control can be put in the hands of users.
The constructivist theory was relied on to inform how users can form their own
knowledge, to be able to use UIs towards “best retrieval results”. UE and Interactivity
attributes were conceptualized in a qualitative and quantitative research methodology.
With user-centric data a measurement model was proposed. The outcome of the
analysis was a measurement model that translated users’ requirements and reported
essential components of an iUI. A personal approach was employed to validate the
users’ requirements following the provisions in the literature (Idoughi et al. 2012).

4. Expected result

Interoperability problem in KOS is also a software engineering problem. Measurable
attributes for UI design that support the interoperability of SEs as KOS is presented
using a measurement model as presented in Figure 2. Cognizance of the advent of
mobile devices that use touch-screens, KOSs are used on the fly. This presents a whole
new twist to the issues raised so far. The eclipse sign shows that more attributes can
therefore be added (found) considering the foregoing dynamics.
5. Conclusion

We conclude that KOS’s success at providing information access and service delivery does not depend solely on the technical aspect of re-engineering KOS’s source and target system for better interoperability alone. For SE to operate with satisfactory results, the purpose of each KOS can be harnessed by a good UI. This implies, therefore that the intellectual aspect, which could encompass statistical modelling (among other approaches) of UI to realize user requirements, is also of paramount importance. This work is descriptive and as such further theoretical insights, empirical research, and testing is still needed.

References


Denis Kos, Sonja Špiranec, Ante Ćović

Mapping perspectival ambiguity in Bioethics: revisiting the viewpoint warrant

1. Introduction

This poster reports on the current state of a doctoral thesis by PhD candidate Denis Kos from the Department of Information and Communication Sciences (University of Zagreb). It presents an original research approach to knowledge organization and provides methodological support in the field of bioethics. Specifically, this thesis aims to confront the problem of perspectival ambiguity in bioethics by considering the notion of perspectives as points of access to diverse outlooks on particular phenomena. The objective of this poster is to elaborate on why ethical issues have potential to be used as a starting point for investigative methodologies that aim to map the perspectival ambiguity surrounding particular (bioethical) phenomena.

2. Theoretical background

The orientation towards perspectival ambiguity calls for different theoretical perspectives stemming both from the workplace context of the candidate and the existing developments that aimed to address similar problems in the past. Hence, the thesis is being based specifically in the conceptualization of integrative bioethics (IB) which approaches bioethical topics with intent to integratively preserve the pluralism of perspectives that contribute to a full characterization of different bioethical phenomena i.e. pluriperspectivistically. The task to define a useful way to organize knowledge for such a complex field is to be supported by the conceptualizations of an interdisciplinary and transdisciplinary knowledge organization (KO) as has been discussed by different authors in the field of KO (Hjørland and Albrechtsen 1995; Broughton 2002; Gnoli 2012; García Gutiérrez 2014; Albrechtsen 2015; Lopez-Huertas 2015; Szostak, Gnoli and Lopez-Huertas 2016). Finally, the candidate considers the recurring evocation of the concept of perspectives in IB and KO (Gnoli 2012; García Gutiérrez 2014; Lopez-Huertas 2015; Szostak, Gnoli and Lopez-Huertas 2016) and the viewpoint warrant in KO (Beghtol 2002) as the third theoretical foundation.

These foundations for the thesis all stem from the critique of the domination of disciplinary perspectives and the idea that we have to address scientific monoperspectivism by representation of how complex our understanding of what we experience can be which warrants the ability to look at knowledge from a number of different perspectives. However, Szostak, Gnoli and Lopez-Huertas (2016) stipulated that in order to achieve this, one first has to determine how much perspectival ambiguity there is in different studied knowledge domains. The candidate recognizes the outcomes
of the thesis as a contribution to understanding this kind of ambiguity with regard to bioethical phenomena and the complex network of perspectives that contribute to their definition.

3. Thesis methodology

The proposed methodology of the accepted thesis proposal consists of:

- a conceptual analysis of the concept of viewpoint warrant in order to elaborate its definition in the context of transdisciplinary KO, and in order to map the role of relevant concepts in its definition like: transdisciplinary knowledge, perspectives, knowledge integration, pluriperspectivism etc.;
- a bibliometrical and content analysis of literature in the field of IB to create a prototype system of relationships between studied bioethical phenomena, explored perspectives and characteristics of documents they are affiliated with;
- a study of expert consensus by using the Delphi method in order to submit the constructed prototype for validation which will be based upon group consensus of its exhaustiveness and appropriateness of term representation.

4. The potential of ethical issues for mapping perspectival ambiguity

The objective of this poster is to focus on the rationale why ethical issues have potential to be used as a starting point for investigative methodologies that aim to map the perspectival ambiguity surrounding particular (bioethical) phenomena. This rationale is based on the idea that the human experience influences what we assume to know and treat as fact. Such objectified truths constellate a personal or group understanding of the world. We propose that these constellations or viewpoints, derived from experience and personal or group history, can be identified when dealing with complex ethical questions. It is the nature of what is usually named an ethical issue to be problematic and because of this to be subject to diverse reasoning on how to define, interpret and resolve such issues. This diversity is the basis of ambiguity and since it stems from objectified constellations of personal or group understandings i.e. perspectives, it is a perspectival ambiguity.

However, ambiguity is often defined as “inexactness” and “the quality of being open to more than one interpretation” (Oxford Living Dictionaries) which is seen as a negative quality. The proposed thesis treats ambiguity as a neutral characteristic, the nature of things found in contested and negotiated understandings of the world. In his paper from 2011, Claudio Gnoli, referring to Beghtol, claims that the demand hidden in the viewpoint warrant, constitutes an ethical principle because no single perspective should be given priority over alternative perspectives. We believe that this demand for social justice also treats ambiguity as a neutral characteristic and as a real encountered quality. In other words, only when the diversity of outlooks is recognized not as problematic or unproblematic, but as an actual condition of social life can ambiguity
become a valid basis of knowledge organization. The notion of integrative bioethics which aims to integratively preserve the pluralism of perspectives that contribute to a full characterization of different bioethical phenomena (i.e. pluriperspectivistically) treats perspectives in the same way. Here goals of IB match those of transdisciplinary KO, since the dialogic interaction of perspectives on ethical issues can only be achieved if we have knowledge about and appreciation of the ambiguity which surrounds them.

5. Conclusion

This poster proposes that ethical issues, as one of the most ambiguous areas of study, are central to studies of perspectival ambiguity which remains a key challenge of transdisciplinary KO, and proposes methodologies to approach the resolution of these challenges. Studies that aim to describe the perspectival nature of documents and the characteristics of found perspectives show just how complex our understanding of what we experience can be and restitutes ambiguity as a valid research interest.

References


María J. López-Huertas, Diego Martín Oliva

A Model for decision making (DM) in territory management: social implications

Introduction
Land management is often a challenge because many factors must be considered in order to take the best possible decision when a problem has to be faced. It is a delicate issue because equilibrium between nature and societal demands need to be sought. It means that the preservation of the territory is tied to the relation and influences of humans in it. Thus, in our view, the concept of territory includes both parts without distinction. Decisions are very often taken without the assistance of knowledge-based systems that guarantee the best objective decision amongst a set of possible choices. On the other hand, systems designed to do such tasks are scarce and normally ignored in the decision-making process.

Objectives
The aim of this research is to elaborate a model on which a knowledge organization system for decision making can be constructed. In this instance, it is a theoretical approach, though with the intention of being applied it in a real setting in subsequent research.

Methodology
In order to build the proposed system, it is very important to have a deep, exhaustive knowledge of the topic at hand. Therefore, a given methodology for territory description and representation will be followed, which will subsequently enable the construction of a knowledge organization system (KOS) that can assist in the process of decision making regarding territory. Territory is a transdisciplinary issue whose study requires a methodological approach that can address such complexity potentially including its societal relationships. Due to this complexity, an eclectic methodology composed of four methods has been developed in order to create a model for a system for decision making regarding territory: 1) The transdisciplinary method of research, according to Nowotny (2006) and Gibbons (2002). This affects how topic is focused and studied, the way in which the transdisciplinary combination of research and action is created and the means of getting a transdisciplinary representation of the knowledge of the territory that will be the foundation for structuring the alternatives for decision making. 2) The use of the Analytical Hierarchical Process (AHP) (Saaty 1980) during the process; this enables a hierarchical order of alternatives for decision making. It has been a fundamental approach because it allows the possible alternatives for a decision
to be articulated. It is a mechanism that enables the choice of best option to be made by pointing to the best possible alternatives in a given decision, based on knowledge organization (KO) of the territory and its context. It can be understood as a technique about preferences with hierarchical aggregation that allows multi-criteria, multi-environment and multi-actor solutions to be found, as well as highlighting the uncertainty inherent in any decision process. It was also used in the selection of specialists in the transdisciplinary group. 3) The construction of the transdisciplinary working group, which was to be established according to the concepts of transversality and participation, including societal actors (Font and Subirats 2000). This includes four possibilities that represent the combination of science, technology and society very well: academics, experts in territory (selected by AHP), computer scientists, information scientists (staff hired) and societal representatives (random selections from census of the territory). 4) A method enabling an understanding of the territory by using knowledge organization of the landscapes that compose it. It is based on the Integrated Evaluation Procedure (Gómez and Riesco 2010), supported by a knowledge base (KB).

Results

After applying the above-mentioned methodology, we arrived at the following main results: 1) To form a transdisciplinary working group, made up of four sets of people:

Table 1. Components and tasks of the Transdisciplinary Group

<table>
<thead>
<tr>
<th>Groups</th>
<th>Tasks</th>
<th>Required Competencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Societal representatives</td>
<td>To carry out the analysis of the territory (methodology). To contribute to the assessment of the alternatives related to the criteria and objectives arising from the analysis. To contribute experiences and feelings about the territory.</td>
<td>A random selection from the census as stated in the methodology.</td>
</tr>
<tr>
<td>Academics Experts</td>
<td>To carry out the analysis of the territory (methodology). To contribute to the assessment of the alternatives related to the criteria and objectives arising from the analysis. To explain the scientific knowledge and to articulate communication within the group.</td>
<td>Geographers, sociologists, economists, lawyers, urban planners, archeologists, art historians, engineers.</td>
</tr>
<tr>
<td>Computer Sci’s, and Tech’s</td>
<td>To support the production of knowledge arising from analysis of the territory. To construct and develop a knowledge base of the system. To perform a geo-referential study of the area connected to the knowledge base.</td>
<td>GIS technicians. Experts in knowledge computation.</td>
</tr>
<tr>
<td>Inform. Sci’s, and Comm. Specialists</td>
<td>To take care of the coordination, information resources, communications and construction of the information system.</td>
<td>Scientific communication professionals, experts in knowledge organization.</td>
</tr>
</tbody>
</table>

The elements of the group should be working as a team, and find a common language based on mutual learning. Thus, it is possible to transfer knowledge in the group.
allowing the production of new knowledge in order to arrive at a decision-making solution. 2) We arrive at the design of a model for constructing a knowledge organization system (KOS) for decision making (Figure 1).

![Model on which the system for decision making is based](image)

Based on a KB as a KOS territory. Once the scene of a given DM is raised and landscapes are identified, the system will propose criteria and objectives. Judged using AHP, it finally comes up with a range of alternatives to be taken the decision.

**Conclusions**

1) A model for the design and construction of a system for DM on territory has been reached. 2) The configuration of a transdisciplinary group could be made by identifying its components and tasks. 3) The foundation of the research on the transdisciplinary method has been proved very effective and has responded to the aims of this study.

**References**


Jaqueline Santos Barradas, Eliane Cristina Maceió Ferreira

State of the art of organization and administration of libraries in Brazil: preliminary results

1. Introduction

This paper presents partial results of the research project in progress called *Organization and Administration of Libraries: state of the art in Brazil*, carried out by professors and students undergoing scientific initiation from the School of Librarianship of the Federal University of the State of Rio de Janeiro (UNIRIO).

The research investigates the development and growth of the area of organization and administration of libraries (OAB), a compulsory curricular component of the curricular matrices of the courses of Bachelor in Librarianship and Degree in Librarianship in Brazil.

The questions that guide the research are: what new subjects are being researched, published and cited in the area of organization and administration of libraries and information units in Brazil? What development and growth are evident in the area?

A state of the art can add significantly to the constitution of the theoretical field of an area of knowledge since it seeks to identify the key contributions to the construction of theory and pedagogical practice. It points out constraints in the field in which research occurs, as well as gaps in dissemination; identifies innovative experiences that seek alternative solutions to practical problems; and also recognizes research contributions in the constitution of proposals in the area (Romanowski and Ens 2006).

2. Objective

The current research aims to reveal the development and growth of the area of organization and administration of libraries from the mapping of scientific production in Brazil, published from 2000 to 2017. For this, it intends to analyze the national and foreign bibliography used in the teaching of themes; inventory scientific output published in formal publication channels; identify research groups registered at CNPq (National Council for Scientific and Technological Development) and their academic-scientific productions; and examine recurring and emerging issues during this period.

3. Methodological procedures

This is a qualitative-quantitative research; qualitative, because it subjectively analyzes the relations found in the research; and quantitative, since it statistically considers the published scientific production per year, in different formal channels of communication.

Initially it was necessary to identify all the courses of librarianship in Brazil through
the e-MEC information platform, via its “interactive consultation” tab. This is an official and unique database of information related to higher education institutions and their respective undergraduate courses organized by states of the Brazilian federation (E-Mec 2017). This procedure gave rise to a new search for information on each course in the electronic portal of the higher education institutions in which they are offered. Until now, it has been possible to collect the programs from the corresponding OAB and the nomenclature of their curricular components and, thus, to identify the basic and complementary bibliography indicated in these programs.

4. Preliminary results

Some difficulties have been encountered so far. The majority of these refer to the collection of OAB curricular component programs, since many of them were not available on websites and portals of Brazilian public universities.

It was therefore necessary to complement the research strategy by: sending e-mails to course coordinators, telephoning the secretariat and course coordination authorities, and contacting the Contact Us resource available on institutional websites. In this way, it was possible to retrieve much of the necessary documentation to enable the later stage, characterized by the analysis of the results. This demanded a much longer time than initially expected, causing a delay in the planned schedule of activities.

Preliminary results indicate the existence of 49 undergraduate courses in librarianship and a degree in the area, offered by 48 higher education institutions. Of these, 31 are offered by public universities and 12 by private universities. Five are offered in the distance learning mode. In Brazil, only four out of the 27 states in the federation do not offer such a course, these four being located in the North region: Acre, Amapá, Roraima and Tocantins. UNIRIO is the only university that offers a degree in librarianship in the country.

In the essential and complementary bibliographies outlined in the curricular programs for subjects constituting librarianship courses, the majority refer to works published at the end of the last century. Several of the course programs consulted indicate some titles by Brazilian authors and researchers, a phenomenon that is also repeated for classic international authors of Administration.

The next stage of the research will analyze the scientific production published in formal, printed and electronic communication channels of the Brazilian Congresses of Library Science, Documentation and Information Science (CBBD) since 2000 (nine in total), by thematic areas.

Regarding research in scientific journals, those selected were categorized in the upper stratum by Qualis/CAPES (A1, A2 and B1) in the thematic area of “Communication and Information”, totaling 16 scientific journals published in Brazil, since Brazilian academic journals do not have a significant impact factor. Their journals
are listed following a classification given by a governmental education agency called CAPES. The analysis of published scientific production is in progress, and has not yet presented results that can be methodologically communicated.

5. Partial considerations

One difficulty that can be pointed out is the fact that private higher education institutions are not willing to share information about their courses. The institutions who collaborated, requested confidential treatment of their data, a determining factor for their exclusion in the cross-section of research. On the other hand, the higher education institutions that shared their non-public data show interest in knowing the results of the research.

On completion, the research is intended to reveal the scientific productivity and the state of art of the area of OAB in Brazil, constituting an up-to-date theoretical-conceptual framework that will allow direct application in classroom teaching, as well as serving as a support for political-pedagogy in librarianship courses, among other outcomes envisaged by the researchers.

References


Moisés Rockembach, Lisiane Braga Ferreira

Web archiving of elections and Brazilian possibilities

Introduction
This research aims to identify some projects developed internationally regarding web archiving in the area of elections, analyzing the importance of these initiatives for the preservation of political and social virtual memory, as well as the applicability of such projects in the context of Brazil.

The analysis follows an exploratory and descriptive approach, from the web archiving projects of elections in the USA, Canada, United Kingdom and France. The methodology used was a bibliographical and documentary study, involving a search of academic papers about web archiving and other related knowledge areas. The sources of the research did not focus on a single institution and its respective webpages but considered other publications that describe relevant projects - thus being analyzed with complementary information whenever possible. In this qualitative approach, characteristics of these projects were collected and analyzed, especially regarding web archiving policies, technologies used and permissions to collect and make information available to users.

Main results and conclusions
Differences were observed among access to information collected, according to the country and its respective legislation. During the 2005 and 2010 elections, the UK requested permission to collect information, requiring curators to contact site owners about filing it before it could be collected and stored. Website owners who declined permission, did not respond or were not located, did not have their sites archived. The 2015 elections were collected after the introduction of new legal deposit regulations in 2013, which allows any UK site to be collected without permission (British Library 2017).

The French web archive, developed by the National Library of France, does not provide the pages collected online, because it understands that the content is protected by copyright and is subjected to legal deposit legislation. This means that access to the web pages collected by the National Library can only be done in reading rooms, with authorized computers (Bibliothèque Nationale de France 2017). Nevertheless, the 2017 French elections were collected by Stanford University Libraries (USA) through the Archive-It platform (Archive-It 2017), which provides online access to this information.

In the United States, web archiving has been carried out weekly by the Congress Library, during electoral periods, since 2000. This web archive documents the sites
associated with presidential, parliamentary and governmental elections. The archived campaign sites of the United States elections also include social networks, in order to get a fuller representation of how candidates were presented to the electorate via the Internet. These collections are comprised from websites of political parties, governments, advocacy groups, bloggers, and other individuals and groups that express relevant views (Library of Congress 2017).

In 2004, the Library and Archives of Canada Act authorized the Library and Archives Canada (LAC) to collect and preserve a representative sample of Canadian web content. LAC has been collecting federal and non-federal content since 2005. (Library and Archives Canada 2017). In the Canadian federal elections of 2015, tweet-gathering discussions emphasized the importance of archiving from the perspective of a historical record not dominated by conventional media. Thus, the collection of this material underscores the need for its ethical use by researchers. The credit of these reference sources in research should be seen as collaborative and information sharing processes, since the voices in the platform are multiple and independent (Ruest and Milligan 2016). In relation to Twitter, some initiatives file the tweets as a broader way of composing the historical moment of the elections along with the memory of the various voices of voters and media coexisting at that time. Despite the difficulties in archiving social networks, especially regarding the closed characteristics of these platforms for commercial reasons, such as with Facebook, it is still considered important to preserve memory from these sources. This possibility of consent is given by the legal permission coming from the contract and developer policy, made available by Twitter, in which this contract allows the capture of only public tweets (Ruest and Milligan, 2016).

The possibilities of web archiving in Brazil are still only emerging incipiently (Rockembach 2018), either through some exploratory research, or by independent initiatives, but have not yet been institutionalized by the Brazilian government in a systematic way. One of the challenges is centered on the legislation about web produced information in Brazil and the corresponding copyright, which is not yet clear, thus making it difficult to achieve a massive archiving of the Brazilian web. However, understanding that the information produced in the Brazilian political elections context should be guided by transparency and publicity, since it is public information, it is possible to understand that the collection and subsequent availability will not cause the same problems that can arise with information from people or private companies published on the web. This proprietary information generates claims possibilities, depending on the interpretation of the current copyright law. Accordingly, it is concluded that Brazil's 2018 general election seems to be an opportunity to test possibilities and web archiving models within the context of Brazil and develop further research.
The web has become a daily and ubiquitous source of information in the lives of many people around the world but, as Brügger and Schroeder (2017) argue, what does this say about historical and social change? Web archiving, then, consists in deliberate and purposeful documentation, with the necessary awareness that the material is being preserved and why it is being preserved, thus, giving many possibilities to researchers from different fields of knowledge. Based on observations and international initiatives, it is possible to justify the research of this digital memory constitution of elections and its multiple uses by various sciences, as well as understanding the methodologies and best practices related to web archiving policies and related technologies. This possibility of preservation and recovery of sites can guarantee the organic, unique, original and probative aspects of the information collected.

References


Workshops
The Politics of classification

Workshop Abstract
Classification schemes are powerful constructions in our information environments. They represent the intentional and unintended biases of those who construct them and are structurally bound by the technical systems in which they are built. As the study of what concepts exist in the world and how expressions of these concepts interconnect, classifications are fundamental to how we come to imagine and subsequently position ourselves embedded within a system of entities in the world. On a more practical level, classification systems become the means by which we materially structure information within knowledge delivery systems, especially in the digital realm.

Given the expressive qualities of classifications, they are also unavoidably political entities: they include and exclude, express and repress, facilitate and restrict. Actively identifying and addressing the political aspects of KO systems is a necessary activity within the KO community.

Workshop description
This workshop begins from a simple premise – classification is political – and asks a simple question: What might we do about it?

Workshop speakers and participants will discuss how we might engage with the politics of classification, as individual researchers and teachers, and as a scholarly community. Our conversation will focus on these questions:

- What grand challenges might we articulate to focus and inspire our work?
- What specific research questions might we be asking?
- What kinds of classes can we offer we be teaching?
- Which disciplines and other areas of study should we more actively engage with?
- Where should we publish to achieve the widest impact?

To provide a framework for this discussion, the workshop will begin with a series of brief presentations by the five conveners. These presentations will provide examples of research that addresses the political through different topics, perspectives, and approaches. Our discussion will then consider the broad themes that emerge from these presentations. Can these themes lead us toward a unifying research vision?

Workshop schedule

- Session 1: 10:00-11:00: 5 Convener presentations, 12-minutes each
- Session 2: 11:00-11:45: Group Breakout and Discussion
- Session 3: 11:45-12:30: Synthetization and Next Steps
Session1: Convener presentations (12 minutes each, 60 minutes total)
This workshop will address the political aspects of classifications through the following 12-minute paper presentations, which will help guide structured discussions.

Meaning through many hands: the politics of classification implementation, Melanie Feinberg
When we say that classification is political, we often focus on the design of classification schemes. But the implementation of any classification continues its design, as meaning piles up through the many anonymous hands that implement classifications, quietly producing the variegated character of datasets. Whatever we do with data – collect it, use it, aggregate it – we’re always making it, hands upon hands upon hands. This presentation focuses on tiny, mundane implementation decisions as an important mechanism of politics through classification.

name / change: Towards a politics of classification, Jonathan Furner
The products of classificatory acts have both instrumental and intrinsic value, to greater or lesser extents. One approach to developing a politics of classification is to try to understand how bad classification happens, to take its effects seriously, and to advocate for remedial change. Focusing on the DDC’s treatment of works about North American Indigenous peoples as a case study, the speaker will present a typology of kinds of, and kinds of potential responses to, classificatory injustice, as a suggested framework for audience discussion of the political dimensions of knowledge organization.

Mechanisms of cultural bias in classificatory activity, Gregory H. Leazer
Given that classifications express cultural and political viewpoints, how can we compare one classification to another to locate bias? What are the mechanisms by which bias is expressed in a classification?

The radiant conflicts of conceptual reconciliation: the politics of biodiversity classification, Robert D. Montoya
A major hurdle in collecting massive amounts of distributed biodiversity data is attempting to reconcile the semantic and taxonomic variability between disparate classification structures. This presentation will focus on the kinds of semantic and conceptual reconciliations that take place in aggregative biodiversity taxonomic work and illustrate the political implications of this “global” work for “local” communities and scientific groups.
Expressing and obscuring throughout: politics, values, and semantics derived from a conceptual analysis of classification terms, Joseph T. Tennis

Given that classification schemes change throughout time one major question is how do we know a change when we see one? Further, those changes might be political, value-laden, or semantic. This talk presents a case for a kind of conceptual analysis that will aid in uncovering characteristics of classes which would allow us to make clear assertions about the nature of change in schemes.

Session 2: Group breakout and discussion (45 minutes)

With the presentations as the point of departure, group attendees will then break off into groups, each facilitated by a workshop convener, who will coordinate recording of the discussion.

Questions of concern for break-out sessions could include:
- What kinds of broader issues and research questions do these presentations as a whole elicit for classification and KO in general?
- Are there any provocative studies that arise from juxtaposing these ideas?
- What kinds of activities can we facilitate at our institutions to incorporate these identified issues at the local level?
- What ancillary domains outside of politics need to be addressed to fully investigate the challenges examined?
- What elements would you like to see in the design of a grand challenge in this area?

Session 3: Synthetization and next steps: group discussion (45 minutes)

Breakout groups will then report their findings to all participants for general discussion. The main aim of this discussion is to synthesize and identify broad themes and research questions that can be used as a basis for future research, workshops, and collaborative activities.

Participants will then discuss ways to coordinate and maintain momentum with regard to these identified topics; possible ways forward may include:
- Providing an ongoing event for research exchange and collaboration that emphasizes critical approaches knowledge organization, inclusive of topics related to politics, diversity, and other cultural approaches to the discipline;
- Generating a community of scholars coalesced around similar interests and aims;
- Identification of conferences and publication outlets for this research, as well as teaching resources;
- Establishing how themes discussed might be integrated into pedagogy;
- And finally, given the presentations of the day, what might be a suitable theme for the next gathering in 2019? Perhaps related to the broader challenges
identified in group discussions.

- **Workshop Outcomes**
  - Identify productive research and teaching directions in the arena of politics as they relate to the teaching of knowledge organization and classification;
  - Establish an ongoing working group on politics, diversity, and classification, including a list of possible topics for the next gathering, tentatively scheduled for 2019.

**Note to participants**

If possible, please bring a laptop or other note-taking device as we will take collective notes on our discussions in an open forum. Panel organizers will create a publicly-accessible word processing document where participants notes can be contributed and collected.
1. Introduction

Numerous studies in the field of knowledge organization have focused on the need to give attention to the different types of informational supports given to photographic documents in order to preserve their information and consequently how they relate to the memory of a place or event. These, in turn, are incorporated in the narratives of academic studies and historical research into places and facts, as well as enabling knowledge about the traditions of a city, an organization or even the social organization of a nation.

Accordingly, treatment and preservation are urgent factors, since the treatment of photographic documents both physically (preservation, adequate packaging, etc.) and intellectually (treatment in relation to content, retrieval and user) is equally important for the retrieval of their information. This recovery embraces concepts leading to an in-depth knowledge of the essentials of descriptive and thematic areas of representation. In a way, the deepening of the foundations demonstrates the role that knowledge organization fulfills: working directly with information and studying the processes to which this information is submitted until it reaches its recipients, considering the elaboration of conceptual models with recovery purposes. Therefore, the treatment of the (intrinsic or extrinsic) content of documents using a descriptive process, with thematic metadata, directly influences users’ retrieval and assimilation of photographic information. Descriptive and thematic treatment are thus complex activities that require careful reflection and specific policies specifying precisely how photographs should be treated, given that this type of document cannot be treated separately in a collection, but has its own characteristics and specificities.

As such, this workshop seeks to conduct a discussion about the fundamentals of representation and organization of information and knowledge and how they are used in archives, libraries, museums and documentation centers. Therefore, thematic representation and metadata are highlighted in the discussion, since they are key processes that directly affect the organization, reception and retrieval of these documents.

2. Considerations on the photographic document in the light of thematic and descriptive representation

The terms "descriptive representation" and "thematic representation" are used in order to highlight the different treatment phases to which the various types of informational resources, including photographic documents, are submitted. This
differentiation refers to the stages that are, at the same time, distinct in their elaboration, but complementary in the sense of building records for catalogs and databases. The descriptive representation seeks the individualization of the document so that its representation is specific; that is, it gathers items by their similarity and establishes relations between them (Mey 1995).

Cataloging should “identify users’ possible information needs, followed by the creation of a tag, a text, or a summary, which will intermediate the access, identification and evaluation of the user in relation to the original resource” (Santos 2013, non-paged). Moreover, cataloging facilitates that when performing a search, the user can easily identify, select and retrieve the document or informational resource that has been described, as well as its contents.

Thematic representation, complementary to descriptive activity, helps in the retrieval of documents since it has the function of exposing their contents. It is characterized by the assignment of subjects to documents through the classification, indexing and preparation of abstracts. According to Dias and Naves (2013, page 7), “thematic treatment [...] has a strong subjective load, because, as the name indicates, it aims to characterize the document from the point of view of its content. This is what many will call the subject of the document, but this subject will depend very much on who does the reading.” Moreover, when it comes to representing themes, there is a gap in relation to photographic documents summed up as follows: “With regard to thematic representation, it can be affirmed that there is an extensive bibliography that consolidates this stage for graphic materials, but not with regard to materials so that theoretical and methodological gaps are perceived that need attention” (Maimone, Gracioso 2007, p. 1).

Thus, both descriptive and thematic representations are integrated characterizing the socialization of the whole collection with the user, and insofar as an organization is elaborated by means of title, author, year and subject, all documents are made available with clarity and objectivity, representations being realized with the highest possible specificity. For the representations to be made, it is necessary, first, to define the typology of the resource worked in order to emphasize the specificities of the materials that will be treated. In relation to the instruments of thematic and descriptive representations, there is a huge lack of manuals, codes and standards that underlie the intrinsic and extrinsic characteristics of the image resource, especially of the photographic document.

Therefore, it is necessary to study the representational essence that this type of document requires, so that for the specialized user, recovery and access to the resource are possible.

In this way, Simionato and Santos (2013, p.3) explain that the recording of the image resource in general, regardless of whether its environment is analog or digital,
demonstrates that the characteristics “ [...] are not addressed and often the attributes are valued only by contextualized textual features”.

Thus, the existing manuals represent the photographic document in the same way as the book. In this context, it is understood that manuals do not provide the necessary support for these specific documents to be properly represented. In addition, manuals do not offer the necessary representation and organization that should inform collections of photographic documents, whether institutional or personal, seeing as they reflect the thought of an era. It is necessary to consider that the forms of organizing and representing documents must be maintained and respected by the institutions which collect them, in order “... not to erase the features of their organicity, as reflected in the manner in which it has been collected, assembled and, even in what was interchangeable in this meeting ...” (Gonçalves, Marcondes 2005, p. 263).

3. Final considerations

It is hoped that the results point to the existence of institutions and principles specific to each field, in relation to how different types of document relates to society, but also show that there are similarities in each of area of study regarding the need for knowledge about the photographic document.

The understanding that all norms in force in different informational environments can encompass singular objects, including photographic documents, is extremely important in terms of constructing a reflection regarding different environments, such as archives, libraries, museums and documentation centers, and regarding the treatment which is carried out by each one, given their particular specificities.

References


List of Contributors and Authors’ Index

Adagunodo, Emmanuel Rotimi 924
Obafemi Awolowo University
Ile-Ife – Nigeria
eadagun@oauife.edu.ng

Adebisi, John Adetunji 720
Obafemi Awolowo University
Ile-Ife – Nigeria
adebisiadetunji@gmail.com

Aderibigbe, Stephen Ojo 924
Lagos State Polytechnic – Nigeria
aderibigbe2000@gmail.com

Ådland, Marit Kristine 684
Oslo and Akershus University College
Norway
maritkr@hioa.no

Afolabi, Babajide Samuel 720, 924, 955
Obafemi Awolowo University
Ile-Ife – Nigeria
bafox@oauife.edu.ng

Agustín-Lacruz, Maria Del Carmen 321
University of Zaragoza – Spain
cagustin@unizar.es

Akhibge, Bernard Ijesunor 720, 924, 955
Obafemi Awolowo University
Ile-Ife – Nigeria
benplus1@gmail.com

Akinruli, Luana Carla Martins Campos 848
Federal University of Minas Gerais – Brazil
luanacampos@insod.org

Akinruli, Samuel Ayobami 848
Federal University of Minas Gerais – Brazil
ayobami@insod.org

Albuquerque, Ana Cristina de 112, 979
State University of Londrina – Brazil
albuanaati@yahoo.com.br

Almeida, Maurício Barcellos 86
Federal University of Minas Gerais – Brazil
mba@eci.ufmg.br

Almeida, Patricia de 58
University of Coimbra – Portugal
mebpatricia@gmail.com

Almeida, Tatiana de 191, 483
Federal University of the State of Rio de Janeiro – Brazil
tatiana.almeida@unirio.br

Andrianasolo, Nathanaëla 754
University of Reunión – France
nathanaela.andrianasolo@univ-reunion.fr

Arakaki, Felipe Augusto 606
São Paulo State University – Brazil
fe.arakaki@gmail.com

Aráujo, Paula Carina de 273
São Paulo State University – Brazil
paula.carina.a@gmail.com

Arias-Coello, Alicia 945
Complutense University of Madrid – Spain
arias@ucm.es

Arquero-Avilés, Rosario 921, 945
Complutense University of Madrid – Spain
carquero@ucm.es

Babik, Wiesław 290
Jagiellonian University in Kraków – Poland
w.babik@uj.edu.pl

Balakrishnan, Uma 693
Common Library Network GBV
Göttingen – Germany
Uma.Balakrishnan@gbv.de

Barité, Mario 339
University of the Republic – Uruguay
mario.barite@fie.edu.uy
Barradas, Jaqueline Santos 966
Federal University of the State of Rio de Janeiro – Brazil
jaqueline.barradas@unirio.br

Barros, Camila Monteiro de 137
Federal University of Santa Catarina – Brazil
camila.c.m.b@ufsc.br

Barros, Thiago Henrique Bragato 103
Federal University of Pará – Brazil
thiagobarros@ufpa.br

Bawden, David 17, 248
City University of London – United Kingdom
d.bawden@city.ac.uk

Beek, Wouter 521
Vrije University Amsterdam
w.g.j.beek@vu.nl

Ben Smida, Kaouther 931
University of Jendouba – Tunisia
kaouther.bensmida@gmail.com

Bernaoui, Radia 856
National Veterinary School of Algiers
Algeria
radiabernaoui@yahoo.fr

Berti Junior, Decio Wey 128
Federal University of Minas Gerais – Brazil
deciowbj@gmail.com

Bezerra, Arthur Coelho 762
Brazilian Institute of Information in Science and Technology (IBICT) – Brazil
arthurbezerra@ibict.br

Biagetti, Maria Teresa 501, 892
Sapienza Rome University – Italy
mariateresa.biagetti@uniroma1.it

Bisi, Pedro Henrique Santos 581
São Paulo State University – Brazil
phbisi@gmail.com

Bourret, Christian 871
University of Paris Est Marne-la-Vallée
christian.bourret@u-pem.fr

Bowker, Lynne 265
University of Ottawa – Canada
lbowker@uottawa.ca

Brito, Marcilio de 167
University of Brasília – Brazil
mdbrito@unb.br

Brito, Marcos de 167
University of Brasília – Brazil
marcosbrito.ill@gmail.com

Brozia, Ana Inés 339
University of the Republic – Uruguay
ins.brozia@gmail.com

Café, Lígia Maria Arruda 137
Federal University of Santa Catarina – Brazil
ligia.cafe@ufsc.br

Cagnazzo, Laura 662
University of Strathclyde – United Kingdom
laura.cagnazzo.2017@uni.strath.ac.uk

Campos, Maria Luiza de Almeida 363
Fluminense Federal University – Brazil
maria.almeida@pq.cnpq.br

Cardillo, Elena 702
Italian National Research Council
Institute of Informatics and Telematics
Italy
elena.cardillo@iit.cnr.it

Cardon, Rémi 931
University of Lille – France
remi.cardon@etu.univ-lille3.fr

Cavalcante, Luciane de Fátima Beckman 112
State University of Londrina – Brazil
lucifbc@gmail.com
Chen, Sophy Shu-Jiun 549
Academia Sinica – Taiwan
sophy@sinica.edu.tw

Chen, Ya-Ning 565
Tamkang University – Taiwan
arthur9861@gmail.com

Cheng, Wei-Ning 282
Nanyang Technological University – Singapore
wcheng009@e.ntu.edu.sg

Chevy-Pébayle, Emmanuelle 804
University of Strasbourg – France
chevry@unistra.fr

Chifu, Adrian-Gabriel 754
Aix-Marseille University – France
adrian.chifu@univ-amu.fr

Cobo-Serrano, Silvia 921, 945
Complutense University of Madrid – Spain
s.cobo@ucm.es

Codina-Canet, Maria Adelina 198
University Carlos III of Madrid – Spain
mcodina@db.uc3m.es

Costa, Luzia Sigoli Fernandes 588
Federal University of São Carlos – Brazil
luziasigoli@gmail.com

Čović, Ante 959
University of Zagreb – Croatia
acovic@ffzg.hr

Crippa, Giulia 152, 483
University of São Paulo – Brazil
giuliac@ffclrp.usp.br
giulia.crippa69@gmail.com

Cruz, Tiago Trindade 795
University of Porto – Portugal
geral@tiagocruz-arq.com

Dal' Evedove, Paula Regina 475
Federal University of São Carlos – Brazil
p.dallevedove@gmail.com

David, Amos 95
University of Lorraine – France
amos.david@univ-lorraine.fr

David, Gabriel 432
University of Porto – Portugal
gtd@fe.up.pt

Díaz Ortúño, Pedro Manuel 314
University of Murcia - Spain
diazor@um.es

Eito-Brun, Ricardo 666
University Carlos III of Madrid – Spain
reito@bib.uc3m.es

El Hachani, Mabrouka 416
University of Lyon 3 – France
mabrouka.el-hachani@univ-lyon3.fr

El Hadi, Widad Mustafa 167
University of Lille 3 – France
widad.mustafa@univ-lille3.fr

Ermert, Axel 29
Institute for Museum Research SMB-PK
Germany
axel.ermert@t-online.de

Evangelista, Isadora Victorino 58, 184
São Paulo State University – Brazil
isadora.biblio@marilia.unesp.br

Fagundes, Priscila Basto 927
Federal University of Santa Catarina – Brazil
priscila.bfagundes@gmail.com

Feinberg, Melanie 975
University of North Carolina
Chapel Hill – United States
melanie_feinberg@unc.edu

Ferreira, Eliane Cristina Maceió 966
Federal University of the State of Rio de Janeiro – Brazil
elianemaceio@yahoo.com.br
<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ferreira, Lisiane</td>
<td>Federal University of Rio Grande do Sul – Brazil</td>
<td><a href="mailto:lisianebf@gmail.com">lisianebf@gmail.com</a></td>
</tr>
<tr>
<td>Braga</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmino, Helder Noel</td>
<td>Federal University of Minas Gerais – Brazil</td>
<td><a href="mailto:hfirmino@eci.ufmg.br">hfirmino@eci.ufmg.br</a></td>
</tr>
<tr>
<td>Monteiro</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fortier, Alexandre</td>
<td>McGill University – Canada</td>
<td><a href="mailto:alexandre.fortier@mail.mcgill.ca">alexandre.fortier@mail.mcgill.ca</a></td>
</tr>
<tr>
<td>204, 900</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fournier, Sébastien</td>
<td>Aix-Marseille University – France</td>
<td><a href="mailto:sebastien.fournier@univ-amu.fr">sebastien.fournier@univ-amu.fr</a></td>
</tr>
<tr>
<td>754</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frank, Ingo</td>
<td>Leibniz Institute for East and Southeast European Studies – Germany</td>
<td><a href="mailto:frank@ios-regensburg.de">frank@ios-regensburg.de</a></td>
</tr>
<tr>
<td>948</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fujita, Mariângela</td>
<td>São Paulo State University – Brazil</td>
<td><a href="mailto:fujita@marilia.unesp.br">fujita@marilia.unesp.br</a></td>
</tr>
<tr>
<td>Spotti Lopes</td>
<td>314, 321, 475</td>
<td></td>
</tr>
<tr>
<td>Gomes, Hagar Espanha</td>
<td>National Council for Scientific and Technological Development Brasilia –</td>
<td><a href="mailto:hagarespanhagomes@gmail.com">hagarespanhagomes@gmail.com</a></td>
</tr>
<tr>
<td>363</td>
<td>Brazil</td>
<td></td>
</tr>
<tr>
<td>Gonçalves, Fernanda</td>
<td>University of Porto – Portugal</td>
<td><a href="mailto:mfgoncalves@fe.up.pt">mfgoncalves@fe.up.pt</a></td>
</tr>
<tr>
<td>432</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gracioso, Luciana de</td>
<td>Federal University of São Carlos – Brazil</td>
<td><a href="mailto:luciana@ufscar.br">luciana@ufscar.br</a></td>
</tr>
<tr>
<td>588</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grácio, Maria Claudia</td>
<td>São Paulo State University – Brazil</td>
<td><a href="mailto:cabrini@marilia.unesp.br">cabrini@marilia.unesp.br</a></td>
</tr>
<tr>
<td>Cabrini</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gravar, Natalia</td>
<td>University of Lille – France</td>
<td><a href="mailto:natalia.grabar@univ-lille3.fr">natalia.grabar@univ-lille3.fr</a></td>
</tr>
<tr>
<td>931</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guimarães, José Augusto Chaves</td>
<td>São Paulo State University – Brazil</td>
<td><a href="mailto:guima@marilia.unesp.br">guima@marilia.unesp.br</a></td>
</tr>
<tr>
<td>184, 460</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hajdu Barat, Agnes</td>
<td>Eötvös Loránd University – Hungary</td>
<td><a href="mailto:hajdu@jgypk.u-szeged.hu">hajdu@jgypk.u-szeged.hu</a></td>
</tr>
<tr>
<td>951</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hajibayova, Lala</td>
<td>Kent State University – United States</td>
<td><a href="mailto:lhajibay@kent.edu">lhajibay@kent.edu</a></td>
</tr>
<tr>
<td>144</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hashtroudi, Negin</td>
<td>Shahid Beheshti University – Iran</td>
<td><a href="mailto:negin.shokrzadeh94@gmail.com">negin.shokrzadeh94@gmail.com</a></td>
</tr>
<tr>
<td>212</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hassoun, Mohamed</td>
<td>National School of Library and Information Science – France</td>
<td><a href="mailto:Mohamed.hassoun@enssib.fr">Mohamed.hassoun@enssib.fr</a></td>
</tr>
<tr>
<td>856</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gomes, Daniel Libonati</td>
<td>Federal University of Pará – Brazil</td>
<td><a href="mailto:daniellibonati00@hotmail.com">daniellibonati00@hotmail.com</a></td>
</tr>
<tr>
<td>103</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hauser, Elliott 300
University of North Carolina
Chapel Hill – United States
eah13@live.unc.edu

Hider, Philip 426
Charles Sturt University – Australia
phider@csu.edu.au

Hudon, Michèle 204
University of Montréal – Canada
michele.hudon@umontreal.ca
michelehudon@gmail.com

Iacono, Antonella 892
Biella Public Library – Italy
antonella.iacono@fastwebnet.it

Ibekwe-Sanjuan, Fidelia 754
Aix-Marseille University – France
fidelia.ibekwe-sanjuan@univ-amu.fr

Ibitoye, Ayodeji Olusegun 770
Bowen University Iwo – Nigeria
ibitoye_ayodeji@yahoo.com

Iyer, Hemalata 347
University at Albany
State University of New York – United States
hiyer@albany.edu

Jamoulle, Marc 702
University of Liege – Belgium
marc.jamoulle@gmail.com

Jansen, Eva Hourihan 392
University of Toronto – Canada
eva.jansen@mail.utoronto.ca

Jansson, Ina-Maria 863
Uppsala University – Sweden
ina-maria.jansson@abm.uu.se

Jorente, Maria José Vicentini 786, 820
São Paulo State University – Brazil
mjorente@yahoo.com.br

Juvêncio, Carlos Henrique 240, 356
University of Brasilia – Brazil
carlosjuv@gmail.com

K. S., Raghavan 347
Indian Statistical Institute
Bangalore – India
ksragav@hotmail.com

Kempf, Andreas Oskar 644
ZBW - Leibniz Information Centre for Economics – Germany
a.kempf@zbw.eu

Kergosien, Eric 931
University of Lille – France
eric.kergosien@univ-lille3.fr

Khoo, Christopher S. G. 231, 282
Nanyang Technological University – Singapore
chriskhoo@pmail.ntu.edu.sg

Kleineberg, Michael 400
Humboldt University of Berlin – Germany
michael.kleineberg@hu-berlin.de

Kos, Denis 959
University of Zagreb – Croatia
dkos@ffzg.hr

Laplante, Audrey 137
University of Montréal – Canada
audrey.laplante@umontreal.ca

Larouk, Omar 416
University of Lyon (Enssib) – France
omar.larouk@enssib.fr

Leazer, Gregory H. 66, 975
University of California, Los Angeles
United States
leazer@ucla.edu

Ledl, Andreas 653
University of Basel – Switzerland
andreas.ledl@unibas.ch
<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lee, Deborah Theresa</td>
<td>City University of London – United Kingdom</td>
<td><a href="mailto:deborah.lee.1@city.ac.uk">deborah.lee.1@city.ac.uk</a></td>
</tr>
<tr>
<td>Lee, Wan-Chen</td>
<td>University of Washington – United States</td>
<td><a href="mailto:leew23@uw.edu">leew23@uw.edu</a></td>
</tr>
<tr>
<td>Leite, António José de Bastos</td>
<td>University of Porto – Portugal</td>
<td><a href="mailto:abastosleite@med.up.pt">abastosleite@med.up.pt</a></td>
</tr>
<tr>
<td>Lima, Gercina Ângela de</td>
<td>Federal University of Minas Gerais – Brazil</td>
<td><a href="mailto:glima@eci.ufmg.br">glima@eci.ufmg.br</a></td>
</tr>
<tr>
<td>Lima, Vânia Mara Alves</td>
<td>University of São Paulo – Brazil</td>
<td><a href="mailto:vamal@usp.br">vamal@usp.br</a></td>
</tr>
<tr>
<td>Lopes, António Tavares</td>
<td>University of Coimbra – Portugal</td>
<td><a href="mailto:atlopes@fl.uc.pt">atlopes@fl.uc.pt</a></td>
</tr>
<tr>
<td>López-Huertas, Maria José</td>
<td>University of Granada – Spain</td>
<td><a href="mailto:mjlopez@ugr.es">mjlopez@ugr.es</a></td>
</tr>
<tr>
<td>Lykke, Marianne</td>
<td>Oslo and Akershus University College – Norway</td>
<td><a href="mailto:mlykke@hum.aau.dk">mlykke@hum.aau.dk</a></td>
</tr>
<tr>
<td>Macedo, Douglas Dyllon</td>
<td>Federal University of Santa Catarina – Brazil</td>
<td><a href="mailto:douglas.macedo@ufsc.br">douglas.macedo@ufsc.br</a></td>
</tr>
<tr>
<td>Machado, Luis Miguel</td>
<td>University of Coimbra – Portugal</td>
<td><a href="mailto:luismachado@fe.uc.pt">luismachado@fe.uc.pt</a></td>
</tr>
<tr>
<td>Maculan, Benildes</td>
<td>Federal University of Minas Gerais – Brazil</td>
<td><a href="mailto:benildes@gmail.com">benildes@gmail.com</a></td>
</tr>
<tr>
<td>Marco-Cuenca, Gonzalo</td>
<td>University of Zaragoza – Spain</td>
<td><a href="mailto:gmarco@unizar.es">gmarco@unizar.es</a></td>
</tr>
<tr>
<td>Marcondes, Carlos Henrique</td>
<td>Fluminense Federal University – Brazil</td>
<td><a href="mailto:marcon@vm.uff.br">marcon@vm.uff.br</a></td>
</tr>
<tr>
<td>Mariz, Anna Carla Almeida</td>
<td>Federal University of the State of Rio de Janeiro – Brazil</td>
<td><a href="mailto:annacarla@unirio.br">annacarla@unirio.br</a></td>
</tr>
<tr>
<td>Mariz, Thales Almeida</td>
<td>Federal University of Rio de Janeiro – Brazil</td>
<td><a href="mailto:thalesmariz@yahoo.com.br">thalesmariz@yahoo.com.br</a></td>
</tr>
<tr>
<td>Martínez-Ávila, Daniel</td>
<td>São Paulo State University – Brazil</td>
<td><a href="mailto:dmartinezavila@marilia.unesp.br">dmartinezavila@marilia.unesp.br</a></td>
</tr>
<tr>
<td>Martín Oliva, Diego</td>
<td>Department of Development of Informatics of Unicaja Bank Group – Spain</td>
<td><a href="mailto:dmartino@correo.ugr.es">dmartino@correo.ugr.es</a></td>
</tr>
<tr>
<td>Martins, Maria Fernanda da</td>
<td>University of Porto – Portugal</td>
<td><a href="mailto:mmmartins@letras.up.pt">mmmartins@letras.up.pt</a></td>
</tr>
<tr>
<td>Melo, Raquel Oliveira</td>
<td>Federal University of the State of Rio de Janeiro – Brazil</td>
<td><a href="mailto:raqueloliveiramelo83@gmail.com">raqueloliveiramelo83@gmail.com</a></td>
</tr>
<tr>
<td>Ménard, Elaine</td>
<td>McGill University – Canada</td>
<td><a href="mailto:elaine.menard@mcgill.ca">elaine.menard@mcgill.ca</a></td>
</tr>
<tr>
<td>Montoya, Robert D.</td>
<td>Indiana University, Bloomington – United States</td>
<td><a href="mailto:montoya@indiana.edu">montoya@indiana.edu</a></td>
</tr>
<tr>
<td>Montoya-Mogollón, Juan</td>
<td>São Paulo State University – Brazil</td>
<td><a href="mailto:juan.mogollon@marilia.unesp.br">juan.mogollon@marilia.unesp.br</a></td>
</tr>
<tr>
<td>Morales, Micaela</td>
<td>University of the Republic – Uruguay</td>
<td><a href="mailto:micaela.morales@hotmail.com">micaela.morales@hotmail.com</a></td>
</tr>
</tbody>
</table>
**Moreira**, Fábio Mosso 176  
São Paulo State University – Brazil  
fabiomoreira@tupa.unesp.br

**Moreiro-González**, José Antonio 908  
University Carlos III of Madrid – Spain  
jamore@bib.uc3m.es

**Moura**, Maria Aparecida 442, 830  
Federal University of Minas Gerais – Brazil  
mamoura@ufmg.br

**Mucheroni**, Marcos Luiz 729  
University of São Paulo – Brazil  
mucheroni.marcosl@gmail.com

**Muriel-Torrado**, Enrique 927  
Federal University of Santa Catarina – Brazil  
enrique.muriel@ufsc.br

**Nakano**, Natalia 786  
São Paulo State University – Brazil  
natinakano@gmail.com

**Naumis-Peña**, Catalina 674  
National Autonomous University of Mexico  
cnaumis@gmail.com

**Ndjock**, Nadine 95  
University of Yaoundé II – Cameroon  
fleur_nadinendjock@yahoo.fr

**Novaes**, Magdala de Araujo 702  
Federal University of Pernambuco – Brazil  
magdala.novaes@nutes.ufpe.br

**Ohly**, H. Peter 856  
GESIS Leibniz Institute for the Social Sciences Germany  
peter.ohly@gmx.de

**Olawuyi**, Nkechinyere Joy 955  
Adeyemi College of Education  
Ondo – Nigeria  
olawuyi_joy@yahoo.co.uk

**Oliveira**, Ely Francina Tannuri de 176  
São Paulo State University – Brazil  
etannuri@gmail.com

**Onifade**, Olufade F.W 770  
University of Ibadan Ibadan – Nigeria  
olufadeo@gmail.com

**Ortega**, Cristina Dotta 74  
Federal University of Minas Gerais – Brazil  
ortega@eci.ufmg.br

**Ortiz-Repiso**, Virginia 908  
University Carlos III of Madrid – Spain  
virginia@bib.uc3m.es

**Padua**, Mariana Cantisani 786  
São Paulo State University – Brazil  
mariana.cantisani@gmail.com

**Paletta**, Francisco Carlos 597, 839, 880  
University of São Paulo – Brazil  
fcpaletta@usp.br

**Park**, Ziyoungh 653  
Hansung University Seoul – Korea  
zgpark@hansung.ac.kr

**Pastor Sánchez**, Juan Antonio 511  
University of Murcia – Spain  
pastor@um.es

**Pattuelli**, M. Cristina 571  
Pratt Institute – United States  
mpattuel@pratt.edu

**Pennington**, Diane Rasmussen 662  
University of Strathclyde – United Kingdom  
diane.pennington@strath.ac.uk

**Pestana**, Olivia 224  
University of Porto – Portugal  
opestana@letras.up.pt

**Pinto**, Adilson Luiz 927  
Federal University of Santa Catarina – Brazil  
adilson@cin.ufsc.br

**Qin**, Jian 409  
Syracuse University – United States  
jqin@syr.edu
Rajan, Lilium Niven 256
University of Washington iSchool
United States
rajanl@uw.edu

Rauch, Mirtha 339
University of the Republic – Uruguay
mbrauch@gmail.com

Reis, Daniela Majorie dos 314
São Paulo State University – Brazil
danielamajorie@yahoo.com.br

Resnick, Melissa P. 702
University of Texas – United States
melissa_resnick@alum.rpi.edu

Ribeiro, Claudio Jose Silva 330
Federal University of the State of Rio de Janeiro – Brazil
claudio.ribeiro@unirio.br

Robinson, Lyn 248
City, University of London – United Kingdom
L.Robinson@city.ac.uk

Rockembach, Moisés 812, 969
Federal University of Rio Grande do Sul – Brazil
moises.rockembach@ufrgs.br

Rodrigues, Ana Célia 43
Fluminense Federal University – Brazil
anyrodrigues@yahoo.com.br

Rodrigues, Fernando de Assis 581
São Paulo State University – Brazil
fernando@rodrigues.pro.br

Rodrigues, Georgette Medleg 240, 356
University of Brasília – Brazil
georgete@unb.br

Rodriguez-Bravo, Blanca 58
University of León – Spain
blanca.rodriguez@unileon.es

Rondot, Camille 804
Paris-Sorbonne University – France
camille.rondot@gmail.com

Salaba, Athena 144, 450
Kent State University – United States
asalaba@kent.edu

Saldanha, Gustavo Silva 191, 483
Federal University of the State of Rio de Janeiro – Brazil
gustavosaldanha@ibict.br

Sales, Luana Farias 711
Brazilian Institute of Information in Science and Technology – Brazil
luanasales@ibict.br

Sales, Rodrigo de 460
Federal University of Santa Catarina – Brazil
rodrigo.sales.s@gmail.com

Salim, Kebbab 856
National School of Journalism and Information Sciences – Algeria
kebbabsalim@yahoo.fr

San Segundo, Rosa 198
University Carlos III of Madrid – Spain
rosa.sansegundo@uc3m.es

Sanches, Tatiana 762
University of Lisbon – Portugal
tsanches@fpie.ulisboa.pt

Sánchez-Casabón, Ana I. 490
University of Zaragoza – Spain
asanchez@unizar.es

Sant’Ana, Ricardo César Gonçalves 176, 581
São Paulo State University – Brazil
ricardosantana@marilia.unesp.br

Santos, Cibele de Araújo Camargo Marques dos 613
University of São Paulo – Brazil
cibeleac@usp.br

Santos, Erika Alves dos 729
University of São Paulo – Brazil
erikasantos@usp.br
Santos, Plácida Leopoldina Ventura Amorim da Costa *606*
São Paulo State University – Brazil
placidasantos@gmail.com

Saorín Pérez, Tomás *511*
University of Murcia – Spain
tsp@um.es

Sayão, Luís Fernando *711*
National Commission of Nuclear Energy – Brazil
lsayão@cnen.gov.br

Scharnhorst, Andrea *521*
Royal Netherlands Academy of Arts and Sciences – Netherlands
andrea.scharnhorst@dans.knaw.nl

Shamenek, Frank S. *702*
Consultant - New York – United States
fshamenek@hotmail.com

Silva, Armando Malheiro da *597, 812, 839*
University of Porto – Portugal
malheiro@letras.up.pt

Silva, Diones Ramos da *330*
Federal University of the State of Rio de Janeiro – Brazil
diones.ramos@icict.fiocruz.br

Silva, Edson Marchetti da *120*
Federal Center for Technological Education of Minas Gerais – Brazil
eamarchettisilva@decom.cefetmg.br

Silveira, Leticia Reis da *588*
Federal University of São Carlos – Brazil
silveira.leticia@gmail.com

Silveira, Naira Christofoletti *483*
Federal University of the State of Rio de Janeiro – Brazil
naira.silveira@unirio.br

Silveira, Teresa *880*
University of Porto – Portugal
tjssilveira@gmail.com

Simionato, Ana Carolina *606, 979*
Federal University of São Carlos – Brazil
simionato.ac@gmail.com

Simões, Maria da Graça *58, 86, 184, 588*
University of Coimbra – Portugal
gsimoes@darq.uc.pt

Siso-Calvo, Brenda *921, 945*
Complutense University of Madrid – Spain
msiso@ucm.es

Siwecka, Dorota *633*
University of Wroclaw – Poland
dorota.siwecka@uwr.edu.pl

Smiraglia, Richard P. *521, 530*
University of Wisconsin, Milwaukee
United States
smiragl@uwm.edu

Soergel, Dagobert *128, 693*
University of Buffalo – United States
dsoergel@buffalo.edu

Soler-Vaquer, Carmen *945*
King Juan Carlos University – Spain
carmen.soler@urjc.es

Soto Hernández, Silvano *674*
IDMS Consulting SC – México
silvano_soto@idms.com.mx

Sousa, Brisa Pozzi de *74*
Federal University of the State of Rio de Janeiro – Brazil
brisapozzi@gmail.com

Souza, Renato Rocha *86*
Federal University of Minas Gerais – Brazil
renato.souza@fgv.br

Špiranec, Sonja *959*
University of Zagreb – Croatia
sspiran@ffzg.hr

Spiteri, Louise F. *557*
Dalhousie University – Canada
Louise.Spiteri@dal.ca
Sun, Guangyuan 231
Nanyang Technological University
Singapore
gsun003@e.ntu.edu.sg

Szostak, Rick 52, 521, 530
University of Alberta – Canada
rszostak@ualberta.ca

Tanti, Marc 940
Aix-Marseille University – France
mtanti@gmx.fr

Tartarotti, Roberta Cristina Dal' Evedove 475
São Paulo State University – Brazil
roberta_tartarotti@yahoo.com.br

Tennis, Joseph T. 273, 308, 450, 975
University of Washington – Unites States
jtennis@uw.edu

Terra, Ana Lúcia Silva 321
Polytechnic Institute of Porto – Portugal
anaterra@iscap.ipp.pt

Tognoli, Natalia Bolfarini 43
Sao Paulo State University – Brazil
nataliabtognoli@yahoo.com.br
natalia.tognoli@pq.cnpq.br

Tramullas, Jesús 490
University of Zaragoza – Spain
tramullas@unizar.es

Tran, Quoc-Tan 624
University of Lille – France
quoc-tan.tran@etu.univ-lille3.fr

Troitiño, Sonia 160, 937
São Paulo State University – Brazil
smtr@marilia.unesp.br

Trombone, Antonella 892
University of Basilicata Potenza – Italy
antonella.trombone@gmail.com

Trzmielewski, Marcin 653
Paul Valéry University Montpellier III
France
marcin.trzmielewski@gmail.com

Vieira, Simone Bastos 167
University of Brasilia – Brazil
sbastosvieira@gmail.com

Voß, Jakob 693
Common Library Network
Verbundzentrale des GBV
Göttingen – Germany
Jakob.Voss@gbv.de

Wybo, Mathilde 931
University of Lille – France
mathilde.wybo@univ-lille3.fr

Yinka-Banjo, Chika O. 770
University of Lagos – Nigeria
chikagog@gmail.com

Zeinolabedini, Mohsen Haji 212
Shahid Beheshti University – Iran
zabedini@gmail.com

Zeng, Marcia Lei 549
Kent State University – United States
mzeng@kent.edu

Žumer, Maja 167
University of Ljubljana - Slovenia
Maja.Zumer@ff.uni-lj.si
Programme Committee

Abdelkrimo Meziane, Research Centre on Scientific and Technical Information, Algeria
Alan Gilchrist, The Cura Consortium, United Kingdom
Amos David, University of Lorraine, France
Ana Isabel Veloso, University of Aveiro, Portugal
Ana Lúcia Terra, Polytechnic Institute of Porto, Portugal
António Lucas Soares, University of Porto, Portugal
Blanca Rodríguez Bravo, University of Léon, Spain
Carlos Almeida, São Paulo State University, Brazil
Carlos Guardado da Silva, University of Lisbon, Portugal
Catalina Naumis, National Autonomous University of Mexico
Claudio Gnoli, University of Pavia, Italy
Clément Arsenault, University of Montréal, Canada
Crispulo Travieso Rodríguez, University of Salamanca, Spain
Dagobert Soergel, University of Buffalo, United States of America
Elias Sanz-Casado, University Carlos III of Madrid, Spain
Francisco Marco, University of Zaragoza, Spain
Gabriel David, University of Porto, Portugal
Heinz Peter Ohly, GESIS: Leibniz Institute for the Social Sciences, Germany
Jens-Erik Mai, Royal School of Library and Information Science, Denmark
Jesús Tramullas, University of Zaragoza, Spain
Jonathan Furner, University of California, United States of America
José António Frias Montoya, University of Salamanca, Spain
José António Moreiro González, University Carlos III of Madrid, Spain
José Augusto Chaves Guimarães, São Paulo State University, Brazil
Joseph Tennis, University of Washington, United States of America
Maja Žumer, University of Ljubljana, Slovenia
Manuel Pinto, University of Minho, Portugal
Maria Cristina Ribeiro, University of Porto / INESC TEC, Portugal
Maria da Graça Simões, Universidade de Coimbra, Portugal
María del Carmen Agustín Lacruz, University of Zaragoza, Spain
Maria Inês Cordeiro, National Library of Portugal
María López-Huertas, University of Granada, Spain
Maria Luisa Alvite Diez, University of Léon, Spain
Maria Manuel Borges, University of Coimbra, Portugal
Mariângela Fujita, São Paulo State University, Brazil
Mario Barité, University of the Republic, Uruguay
Michèle Hudon, University of Montréal, Canada
Moisés Martins, University of Minho, Portugal
Óscar Mealha, University of Aveiro, Portugal
Radia Bernaoui, Higher National School of Veterinary, Algeria
Rebecca Green, The Library of Congress, United States of America
Renato Rocha Souza, Federal University of Minas Gerais / Getúlio Vargas Foundation
Ricardo Eito-Brun, University Carlos III of Madrid, Spain
Richard Smiraglia, University of Wisconsin-Milwaukee, United States of America
Rick Szostak, University of Alberta, Canada
Rosa San Segundo, University Carlos III of Madrid, Spain
Stella Dextre Clarke, Independent consultant, United Kingdom
Thomas Dousa, The University of Chicago Library, United States of America
Vera Dodebei, Federal University of the State of Rio de Janeiro, Brazil
Widad Hadi, University of Lille, France
Wieslaw Babik, Jagiellonian University in Kraków, Poland