Knowledge Organization at the Interface

Proceedings of the Sixteenth International ISKO Conference, 2020
Aalborg, Denmark

Organized by International Society for Knowledge Organization (ISKO), Marianne Lykke, Tanja Svarre, Mette Skov and Daniel Martínez-Ávila

Edited by Marianne Lykke Tanja Svarre Mette Skov Daniel Martínez-Ávila
Knowledge Organization
at the Interface
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.d-nb.de.
# Table of Contents

## Introduction

*Giovanna Aracri, Assunta Caruso, Antonietta Folino: An Ontological Model for Semantic Interoperability Within an Earth Observation Knowledge Base*  
11

*Webert Júnior Araújo, Gercina Ângela de Lima: A Methodological Proposal Towards Domain Ontology Enrichment*  
23

*Mario Barité, Mirtha Rauch: Cultural Warrant: Old and New Sights from Knowledge Organization*  
31

*Maria Teresa Biagetti: Bibliographical Relationships in Knowledge Organization Systems: A Historical-Theoretical Perspective*  
41

*Ceri Binding, Claudio Gnoli, Gabriele Merli, Marcin Trzmielewski, Paul-Valéry, Douglas Tudhope: Integrative Levels Classification as a Networked KOS: A SKOS Representation of ILC2*  
49

*Pino Buizza: Thesaurus and Heading Lists: Equivalences and Divergences*  
59

*D. Grant Campbell, Alex Mayhew: Inheritance and Lamination in the Representation of Bibliographic Relationships*  
69

*Josir Cardoso Gomes, Marco André Feldman Schneider: Ethical Perspective on Classifications of Religions: The Protestant Rise in Brazil*  
78

*Yi-Yun Cheng, Khanh Linh Hoang, Bertram Ludäscher: Cacao, Cocoa, or Cocoa?: Reconciliation of Taxonomic Names in Biodiversity Heritage Library*  
88

*Stephanie Colombo: Representation and Misrepresentation in Knowledge Organization: The Cases of Bias*  
98

*Giulia Crippa, Andre Vieira de Freitas Araujo: Order of Knowledge, Selection and Bibliographical Tension in the 16th Century: Between Gesnerian Universality and Possevian Anti-Heretism*  
105

*Amelie Dorn, Renato Rocha Souza, Enric Senabre, Thomas Palfinger, Eveline Wandl-Vogt, Barbara Piringer: Crafting a System for Knowledge Discovery and Organisation: A Case-Study on KOS for a Non-Standard German Legacy Dataset*  
115

*Sharon Farnel, Ali Shiri: Indigenous Community Driven Knowledge Organization at the Interface: The Case of the Inuvialuit Digital Library*  
123

© Ergon - ein Verlag in der Nomos Verlagsgesellschaft
Amel Fraisse, Samantha Blickhan, Victoria Van Hyning: Towards an Open, Inclusive and Sustainable Knowledge Organization Models

Jonathan Furner: New Formats, Shifting Fortunes: Late-Twentieth-Century KO in the Wild

Francisco-Javier García-Marco, Fernando Galindo, Pilar Lasala, Joaquín López del Ramo: Advancing the Interoperability of the GLAM+ and Cultural Tourism Sectors through KOS: Perspectives and Challenges

Ann M. Graf: Domain Analysis of Graffiti Art Documentation: A Methodological Approach

David Haynes: Understanding Personal Online Risk to Individuals Via Ontology Development

Antoine Henry, Widad Mustafà El Hadi: The Use of Community to Organize Knowledge: The Case of an Energy Company

Philip Hider: Fiction Genres in Library Catalogues and Social Cataloguing Sites

Maximilian Hindermann, Andreas Ledl: BARTOC FAST: A Federated Asynchronous Search Tool for Remote Vocabulary Access

Chris Holstrom, Joseph T. Tennis: Visibility, Identity, and Personal Expression: Qualitative Case Studies of Social Tagging on MetaFilter

Gregory H. Leazer, Robert Montoya, Jonathan Furner: Numerical Classification and Complexity: Developing a Classification of Classifications

Deborah Lee, Lyn Robinson, David Bawden: Operatic Knowledge Organisation: An Exploration of the Domain and Bibliographic Interface in the Classification of Opera Subgenres

Daniel Libonati Gomes, Thiago Henrique Bragato Barros: The Bias in Ontologies: An Analysis of the FOAF Ontology

Lucinêia Souza Maia, Gercina Ángela de Lima: A System for Specifying Semantic Relations for Knowledge Representation

Carlos Henrique Marcondes, Célia da Consolação Dias: Representing Faceted Classification in SKOS

Daniel Martínez-Ávila, Fidelia Ibekwe, Fernanda Bochi: The Epistemic Communities and Evolution of Knowledge Domains: A Domain Analysis of the Journal Education for Information

Paul Matthews: Knowledge Organisation Systems for Chatbots and Conversational Agents: A Review of Approaches and an Evaluation of Relative Value-Added for the User

Claire McDonald: Call Us by Our Name(s): Shifting Representations of the Transgender Community in Classificatory Practice
Ádne Meling: A Critique of the Use and Abuse of Typologies in Cultural Policy Analysis

Juan Bernardo Montoya-Mogollón, Sonia Troitiño: Digital Forensics Science and Knowledge Organization: An Interdisciplinary Approach to Addressing the Conceptual Challenges of Born-Digital Records

Katherine Morrison: Committed to a Narrative: Expressions of Knowledge Organization at The Henry Ford Museum of American Innovation

Catalina Naumis-Peña, Hugo Alberto Guadarrama-Sánchez, Luis Enrique Sánchez-Rodríguez, Rosa de Guadalupe Hernández-Villeda: Terminological Relations of a Thesaurus for University Cultural Infrastructure Terms

Inger Beate Nylund: Using the Concept of Warrant in Designing Metadata for Enterprise Search

Lucia Maria Velloso de Oliveira, Bianca Therezinha C. Panisset, José Antonio da Silva: Types of Documents: Representations of Who We Are and How the Government Works

Ziyoung Park, Hosin Lee, Seungchon Kim, Sungjae Park, Dasom Jung, Seunghee Son, Yoonwhan Kim, Hyewon Lee: Organizing Performing Arts Records of Korean Traditional Music as Linked Open Data

Ziyoung Park, So Young Yoon, Seunghee Son, Yoonwhan Kim: Constructing Semantic Periodical Index Database Focusing on the Visegrad Group’s Transition Process (writing problems)

Brigita Perchutkaite, Marianne Lykke: Facilitating University–Industry Interaction by Visually Showcasing Researcher Profiles Via Metadata

Günter Reiner, Philipp Adämmer: Similarities Between Human Structured Subject Indexing and Probabilistic Topic Models

Athena Salaba: Knowledge Organization Requirements in LIS Graduate Programs

Gustavo Saldanha, Giulia Crippa: Concept Theory and Conceit Theory Ontology and Logology Between Conceptuality and Non-Conceptuality in Knowledge Organization


Carlos Guardado da Silva, Luís Corujo, Jorge Revez: The Classification Plan for Local Administration: Portuguese Archives and the Knowledge Organization in Practice

Richard P. Smiraglia, Rick Szostak: Identifying and Classifying the Phenomena of Music
Linda C. Smith: Interdisciplinary Searching as a Use Case for Vocabulary Mapping 428

Rick Szostak, Richard P. Smiraglia, Andrea Scharnhorst, Ronald Siebes, Aída Slavic, Daniel Martínez-Avila, Tobias Renwick: Classifications as Linked Open Data: Challenges and Opportunities 436

Natália Tognoli, Suellen Oliveira Milani, José Augusto Chaves Guimarães, João Batista Ernesto de Moraes: The Subject Dimension of Authorship: A New Perspective of Provenance in KO 446

Uma Balakrishnan, Dagobert Soergel, Olivia Helfer: Representing Concepts through Description Logic Expressions for Knowledge Organization System (KOS) Mapping 455

Mario Barité, Mirtha Rauch: Classification System for Knowledge Organization Literature (CSKOL): Its Update, a Pending Task? 460

Thiago Henrique Bragato Barros: Touching from a Distance: Concept Theory and Archival Hierarchical Classification 465

Amelie Dorn, Yalemisew Abgaz, Gerda Koch, José Luis Preza Díaz: Harvesting Knowledge from Cultural Images with Assorted Technologies: The Example of the ChIA Project 470

Francisco-Javier García-Marco: Knowledge Organization in Historical Information Systems Revisited: Changes in Society, Technology and Expectations 25 Years Later 474

Negin Shokrzadeh Hashtroudi, Mohsen Haji Zeinolabedini: Representing Entities and Characteristics of Iranian Performing Arts Based on IFLA Library Reference Model (IFLA-LRM) 479

Christopher S.G. Khoo, Rebecca Y.P. Kan: An Ontology for Conceptual Analysis of Signature Pedagogies 484

Michael Kleineberg: Classifying Perspectives: Expressing Levels of Knowing in the Integrative Levels Classification 489

Wan-Chen Lee: Linking, Mapping, Matching, and Change: Contemporary Use of Ranganathan’s Three Planes of Work in Classification Activity 494

Xiaoyue Ma, Pengzhen Xue, Nada Matta: Reconstruction of Crisis Knowledge Ontology by Integrating Temporal-Spatial Analysis 499

Luis Machado, Graça Simões, Claudio Gnoli, Renato Souza: Can an Ontologically-Oriented KO Do Without Concepts? 502

Victor Odumuyiwa, Yetunde Zaid, Olatunde Barber: Enhancing Knowledge Organization Through Implicit Collaboration in Crowdsourcing Process 507
Lucia Maria Velloso de Oliveira, Bianca Therezinha C. Panisset, José Antonio da Silva: Mediation in Archives: Organization, Classification and Transparency

Olivia Pestana, Rui Sousa-Silva: Knowledge Organization in the New Era Using DIY Corpora as Writing Assistants

Marcos Gonçalves Ramos, Priscila Ramos Carvalho, Rosali Fernandez de Souza: Amazônia and Amazon: Domain Analysis with Iramuteq in Scopus and LISA Databases

Tobias Renwick, Rick Szostak: A Thesaural Interface for the Basic Concepts Classification

Ana Lúcia Terra, Maria Del Carmen Agustín-Lacruz, Mariângela Spotti Lopes Fujita: The Role of Knowledge Organization in Scientific Communication: An Overview on JCR's Psychology Journals Guidelines about Title, Abstract and Keywords

Julietti de Andrade, Marilda Lopes Ginez de Lara: The social role of knowledge organization in Evidence Based Health

Radia Bernaoui, Dagobert Soergel: Social Network Communication and Effects on Innovation: The Case of the Agrifood Sector in Algeria

Djadeu Nguemedyam Colette: Organization and Sharing of Knowledge on Selective Household Waste Collection for Hygiene and Sanitation in the City of Yaoundé, Cameroon

Rodrigo Aldeia Duarte, Rosali Fernandez de Souza, Gustavo Saldanha: Devising a Concept of User for Archival Science: An Analysis of the Brazilian Scientific Literature

Isadora Victorino Evangelista, Thiago Henrique Bragato Barros: Ethical Aspects in Knowledge Organization: A Discourse Analysis at ISKO International Events

Negin Shokrzadeh Hashtroudi, Mohsen Haji Zeinolabedini: Educational Practices of Knowledge Organization in Iran: A Historical Review

Maria Leticia Pereyra Lanterna, María José López-Huertas Pérez, Francisco José Morales Calatayud: Approach to Domain Community Health and its Implications for Information Management

Bruno Henrique Machado, Rafael Semidão, Telma Campanha De Carvalho Madio, Daniel Martínez-Ávila: Provenance as an Ethical Measure for the Archival Knowledge Organization of Photographs

Alex Mayhew: Phylomemetic Cataloguing: Expanding Bibliographic Relationships Beyond FRBR
Marcos Luiz Cavalcanti de Miranda, Maria Luiza de Almeida Campos: Knowledge Organization Cultural Studies and Their Influence on Knowledge Organization Systems from the Douglas John Foskett’s Perspectives 562

Katerina Lynn Stanton, Rachel Ivy Clarke: The Design Domain is Divided: Issues in Interdisciplinary Library Classification 564


Natália Tognoli, Lucas Correa: Knowledge Organization Systems as Accountability Tools in Archival Science 569

Fernanda Valle, Gustavo Saldanha: Autism Disorder in KO: Classification, Representation and Social Impact 572

Subject Index 575

Author Index 580

International Scientific Committee 582
The 16th International ISKO conference, under the theme Knowledge organization at the interface was planned to take place from the 6. – 8. July 2020, in Aalborg (Denmark) at Aalborg University, Department of Communication and Psychology.

The conference theme explored the connected themes of knowledge organization systems and their role in knowledge organization, knowledge sharing, and information searching. The conference invited contributions with practical solutions as well as the theory behind the design, development and implementation of knowledge organizing systems, ranging from controlled vocabularies, classification systems, metadata schemas through to ontologies and taxonomies.

The conference topics included:

- Knowledge organization across domains, media and technologies
- Knowledge organization as understanding and communication
- Knowledge organization as driver for development and change

Proposals for full papers, short papers, posters, round table discussion, and workshops were welcomed. The conference call invited KO academics, practitioners, developers and students of innovative ideas and solutions to submit abstracts for consideration. Criteria for acceptance included originality, clarity of expression, and relevance to the conference theme. Proposals should have a sound basis in KO theory, be previously unpublished research, and not under review for another conference or journal.

Anonymized full paper, short paper and poster submissions were double-blind reviewed. Submissions for round table discussions and workshops were single-blind reviewed, and should not be anonymized. The review process was carried out as a two-step process where first 1) abstracts were reviewed and accepted for further development, and 2) later accepted as either full paper, short paper, or poster when a full, developed version had been submitted. All accepted papers and posters are published in the conference proceedings.

Accepted papers

During the two-step review process a total of 48 full papers, 17 short papers, and 14 posters were accepted for publication and presentation at the conference. The papers covered a wide range of topics within the conference themes, i.e., knowledge transfer, concepts and conceptualization, fiction genres, ethical aspects, classificatory structures, representation, probabilistic models, social tagging, domain analysis, music classification, legacy data, document types, semantic networks, bibliographic relationships, faceted classification, KOS mapping, warrants, KO education, museum
organization, and archival organization. The papers discussed theoretical issues related to knowledge organization and the design, development and implementation of knowledge organizing systems as well as practical considerations and solutions in the application of knowledge organization theory. The papers covered knowledge organization systems ranging from classification systems, thesauri, metadata schemas through to ontologies and taxonomies. Scholars from 28 countries from all over the world contributed to the conference proceedings. As the conference was cancelled due to COVID-19, no conference program has been developed, and the papers are organized in full papers, short papers, and posters and within this organization alphabetically by first author.

We would like to thank to those who helped make this publication possible. In particular, we are very grateful to the scholars who submitted abstracts and contributed with excellent papers as well as to the reviewers who made their effort to review and suggest improvements to the submissions.

Aalborg, July 14 2020
Marianne Lykke, Tanja Svarre, Mette Skov, and Daniel Martínez Ávila
Abstract:
This paper presents an ontological model which will be included in an EO Knowledge Base (KB) covering four thematic strands. As a multitude of heterogeneous data will be made available through the KB, it is essential to ensure high standards of discoverability, accessibility, and interoperability. The overall aim therefore, is to align and integrate a set of existing semantic resources and ad hoc vocabularies into a single ontological conceptual model, which defines the specific domains and which will facilitate information and knowledge generation from EO data. Thus, guaranteeing semantic interoperability within the KB platform, ensuring harmonised access to and retrieval of the vast volume of data produced, turning it into usable information and knowledge.

1.0 Aim and scope of the study
Integrating complex data, dynamic in nature, from heterogeneous resources, and without broadly applied standards constitutes a real challenge for users trying to make sense of the increasing amount of information made publicly available in any domain (Bodenreider et al. 2002). Earth Observation (EO) data, in particular, has increased considerably over the last decades and end-users are facing many challenges in accessing and analysing this data. In order to address the organisation and homogenisation of the huge volume of information that Earth Observation research is producing nowadays, scientists, and not only, need support from lexical and semantic tools, such as terminologies, vocabularies, nomenclatures, code and synonym sets, lexicons, thesauri, ontologies, taxonomies and classifications (De la Iglesia et al. 2013). Information Science typically defines information in terms of data, knowledge in terms of information, and wisdom in terms of knowledge (Rowley 2007). Generating information and knowledge from data is about understanding and connecting. Some initiatives, such as Copernicus, Group on Earth Observation (GEO), INSPIRE and others 2, focusing on air quality monitoring, atmospheric conditions and pollution emissions, have generated a large volume of data. These data have been collected by using different devices simultaneously and dissimilar modalities, thus access to them remains difficult.

On the basis of the above actions, the ERA-PLANET Programme 3 with the enrolment of the most active experts in the EO field aims to develop, among other objectives, a Knowledge Platform to ensure harmonized access to the vast amount of data produced.

---

1 Although the authors have cooperated in the research work and in writing the paper, they have individually devoted specific attention to the following sections: Aracri: 3.3, 3.4 and 4.0; Caruso: 1.0 and 3.2; Folino: 2.0 and 3.1.


The purpose is to gather distributed data coming from in-situ sensors and satellite-based remote sensing technologies, by means of practices and infrastructures able to organize, interpret and summarize them. To meet this challenge, the Programme is made up of 4 strands, each one corresponding to a project. All the projects try to provide more reliable information to policy makers concerning the status of the Earth in order to discuss and identify common strategies able to limit climate change, which is dangerous both for environmental and human health, by promoting sustainable development. Such data and information need to be made available through an interoperable system for data sharing and management able to ensure data quality and to interpret the meaning of data, turning them into usable information and knowledge.

Existing EO frameworks support a variety of geographic data set types as well as tools for data management, analysis and visualisation, but often they do not provide any mechanism to tackle semantic heterogeneity issues (Fugazza et al. 2010). Even when EO platforms do include a number of vocabularies, if they are not aligned, the retrieval of all the information regarding a certain topic will not be guaranteed. Relating terms from distinct vocabularies creates richer structural information that can be used for improving search and query expansion (Craglia 2011). To fulfil this requirement an ontology is under development and a set of semantic resources are being integrated and aligned through a corpus-based approach. The overarching goal is to organize and to give access to integrate complex data, dynamic in nature, from heterogeneous resources, and without broadly applied standards, in order to improve the ability of end-users to explore and exploit EO data. The fulfillment of these ambitious goals, as well as the necessity to integrate our semantic resource into a comprehensive platform oriented us towards the choice of an ontology since it provides a conceptual framework which is more structured, adaptable and reusable.

2.0 Literature review

In this section we provide a brief literature review regarding methods and projects oriented to align and integrate sets of existing semantic resources and ad hoc vocabularies, regardless of the domain within which they have been applied. As for the alignment, the term used for referring to the establishment of a variable degree of correspondence between concepts that belong to different controlled vocabularies is ontology mapping. The increasing number of ontology matching methods and tools and the necessity of reaching a consensus on their evaluation determined the advent of the Ontology Alignment Evaluation Initiative (OAEI).

---


7 http://oaei.ontologymatching.org/
In general, in order to be aligned and subsequently published as linked open data on the Semantic Web, several existing controlled vocabularies have been previously converted in SKOS format. Some of them have been mapped without the systematic application of automatic approaches or the direct involvement of domain experts. The indexing languages RAMEAU, Library of Congress Subject Headings (LCSH) and Subject Headings Authority File (German: Schlagwortnormdatei SWD) (Landry 2009), as well as the Dewey Decimal Classification (DDC), the Library of Congress Classification (LCC) and the Medical Subject Headings (MeSH) (Vizine-Goetz et al. 2004), the Thesaurus for Economics (STW) and the Thesaurus for the Social Sciences (TSS) (Mayr and Petras 2008) and the Thésaurus du Tourisme et des Loisirs (Caruso and Folino 2015) have been mainly aligned without exploiting automatic approaches.

Some other research works propose the semi-automatic detection of exact matches between concepts, and several matching systems are based on the computation of string similarity measures, rather than on semantic criteria. However, in almost all studies, the evaluation phase is performed manually. As stated in Morshed et al. (2011), some tools, such as S-match, use external resources (i.e. WordNet) as a background for recognising semantic relationships, but this approach seems less suitable for domain-specific terminologies. The approach here explained concerns the alignment between AGROVOC and other six KOSs more or less related to the field of agriculture. For each pair of concepts some string similarity measures are computed and the average value is taken into consideration for the subsequent manual validation. In order to perform this phase, experts have considered the status of the term (preferred or not), the hierarchy, the equivalent labels in other languages and the notes associated to concepts. The use of hierarchy as a disambiguation technique has also been used in presence of one-to-many alignments in Tordai et al. (2009) and, for each concept to be mapped, it takes into consideration both broader and narrower alignments. The authors have adopted a combination of techniques for establishing mappings between concepts coming from thesauri belonging to the cultural heritage domain: syntactic exact match techniques, a linguistic analysis and a technique deploying the ontology structure.

The need for a manual evaluation depends on several problems generating incorrect matches when using automatic approaches. Some of them are listed in Kempf et al. (2014): terms share the same lexical value but their broader and narrower terms or their scope notes are different; terms in different domains seems to be similar but their meaning is different; the matching between a synonym and a preferred term generates an incorrect equivalence.

To sum up, we refer to one of the most complete classifications of matching techniques (Euzenat and Shvaiko 2013), shown in Figure 1.

Furthermore, starting from the assumption that a domain of interest can be represented through a corpus of text documents, it can be assumed that the knowledge domain that should be encoded into an ontology is represented through a domain corpus, and that the evaluation should output some measures that express the coverage and the adequacy of the ontology with respect to said domain (Rospocher et al. 2012). The integration of ad hoc vocabularies should therefore begin with the acquisition of domain-specific terminology (ex. Liddle et al. (2003); Navigli and Velardi (2004);
Methodologies similar to the one used in the present paper are brought forth by Brewster et al. (2004), in which the authors present a method for evaluating an ontology by comparing it with a domain-specific corpus, and by Cui (2010), who compares the coverage, semantic consistency, and agreement of four thematic ontologies by checking them against a corpus of domain literature.

Figure 1. Classification of matching techniques

3.0 Method

3.1 Corpus-based terminology extraction

New domains with specific conventions and new terminology are continuously appearing. The development of terminology lists is a previous step in KOS building that allows the use of automatic or semi-automatic techniques that could facilitate the labour (Peñas et al. 2001). Terminology extraction deals with the identification of terms which are frequently used to refer to the concepts in a specific domain, and therefore most likely representative of it. The main aim of term extraction here was to carry out a corpus-based terminological evaluation of the existing semantic resources in the EO domain in order to assess whether they adequately cover the terminology used in the domain corpus. Terminology extraction has been undertaken using the T2K² (text-to-knowledge) tool (Dell’Orletta et al. 2014), specifically conceived to identify and extract simple and compound terms from unstructured texts. The main assumption on which T2K², along with most terminology extraction software, is based, is that the relevant concepts of a text are conveyed by the terms that will occur most frequently. The tool performs a linguistic analysis of the texts, the result of which consists of a terminological vocabulary accompanied by semantic and conceptual information about the terms themselves, which add to the value of the output. The list of candidate terms has been sorted by frequency and subsequently manually revised. Understanding whether a given KOS adequately covers the domain of interest is a common and important issue when evaluating a semantic resource. After having matched and evaluated the existing resources’ semantic/terminological coverage, which will be detailed in the following section, the terms appropriate to the sub-domains and objectives involved were used to produce/integrate the ontology. Both terms common to the corpus and to the existing vocabularies, as well as terminology unique to the domain corpus, i.e. not present in any of the examined vocabularies, was used to integrate the ontology to be incorporated in the EO Knowledge Base platform.
3.2 Evaluation of EO vocabularies semantic coverage

Alongside the corpus construction, relevant existing terminological resources related to the Environment have been collected to be used as references. To date, thesauri such as the General Multilingual Environmental Thesaurus (GEMET)\(^9\), the EARTH Thesaurus\(^{10}\), the AGROVOC Thesaurus\(^{11}\), along with the INSPIRE Feature Concept Dictionary and Glossary\(^{12}\), have been taken into consideration. These terminologies have been downloaded in an easily computable format and in the form of flat-lists in order to compare all their terms to those extracted in the previous phase.

Over 5,550 terms have been extracted from the GEMET Thesaurus, while around 14,000 terms from the EARTH thesaurus. Meta-terms along with Macro areas have not been taken into consideration. The AGROVOC thesaurus includes approximately 45,500 terms. Almost 200 terms have been extracted from the INSPIRE glossary and 360 from the Feature Concept Dictionary. Considering their partial semantic overlapping, some vocabularies are already mapped to each other in order to allow federated access to information. An initial comparison has been carried out between the four term lists and those extracted from our sub-corpora in order to identify exact matches and understand if the concepts relevant to our specific purposes are present in the existing resources.

The comparison has been carried out through the use of WordSmith Tools\(^{13}\), a software tool that allows the comparison of the wordlists generated from the existing resources and the corpus. For instance, 450 candidate exact matches were found when comparing GEMET with the sub-corpus terminology, 26 exact match candidates in INSPIRE, 886 exact matches between AGROVOC and the specialized term list, and 763 candidate exact matches were automatically identified between the terminology present in EARTH and the terms extracted from one of the sub-corpora. A further manual analysis however, results in the identification of close matches, broader and narrower matches.

\(^{10}\) https://vocabularyserver.com/cnr/ml/earth/en/.
\(^{13}\) https://lexically.net/wordsmith/.
3.3 Definition of an ontological conceptual model

In this section we present the methodological choices and the steps involved in the development of the ontological model covering the thematic strands, which are strictly related but not necessarily overlapping. In a nutshell, an ontology is a complex formal vocabulary\textsuperscript{14} which requires background knowledge of the domain in order to explicit an efficient conceptual framework. Within the ontology the main issue is how concepts are logically implicated and which kind of information/knowledge we would like to get out from concepts and relationships, even if there are no \textit{a-priori} links between them. EO data and information are both extremely heterogeneous and dynamic in terms of formats, meanings, languages, etc., and captured by using different technologies, applications and systems\textsuperscript{15}. For the ERA-PLANET program, the major challenge regards providing decision support tools aimed at guaranteeing access to relevant information, whenever needed and in a comprehensible format. Thus, data and information can be reused and managed for various purposes, depending on the challenge.

The ontology to be included in the ERA-PLANET Knowledge Platform has been developed through the use of Protégé, an open tool used to design and integrate ontologies. The aim is to support decision making in accordance with the Sustainable Development Goals (SDGs) and to monitor activities in evaluating measures to support environmental policies. To provide a complete and detailed representation of the concepts faced in the projects a top down approach has been adopted by implementing a high level conceptualization due to the inclusive nature of the ontology. Hence, at the top level there are the main classes which are key concepts across the domains and the access points to explore and browse the entire conceptual structure without evident boundaries among the sub-domains. Classes are, in turn, organized in further sub-classes

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{Term list comparison}
\end{figure}

\begin{itemize}
\item \textsuperscript{14} https://www.w3.org/standards/semanticweb/ontology.
\end{itemize}
at different hierarchical levels\textsuperscript{16} and are at the same time member of the superclass and root of another subclass. The result is a taxonomy in which concepts are hierarchically organized by means of the specification of generic relationships in a universally acceptable manner. Even if concepts included in the ontology are related to the subdomains, there are those that are strictly domain-based (ex. Ecosystem - Terrestrial_ecosystem - Anthropogenic_terrestrial_ecosystem - Cropland), thus they present a very high level of specialization, while others, whose meaning is less domain-specific, present few hierarchical levels (ex. Dataset - Copernicus scenes). Therefore, at the moment, the same granularity cannot be guaranteed in all classes. Classes are not populated by individuals, but are related to each other by means of direct and inverse properties (ObjectProperties) defined according to the type of relationship that is useful to explicit. The match between classes and properties generates a statement in the form of subject-predicate-object expressions (ex. “Indicator 15.3.1 measures Target 15.1” and vice versa “Target 15.1 isMeasuredBy Indicator 15.3.1”).

Figure 4. Ontological model

3.4 Ontology and vocabulary alignment

The ontology development benefits from the terminology extraction and comparison phases described above. Some concepts have been added to expand the semantic granularity, and in order to better contextualize each concept, some additional information, such as definitions coming from other vocabularies or links to multiple sources, has been included. In defining mappings a semi-automatic approach has been adopted and a unidirectional mapping has been implemented from our ontology to the above mentioned controlled vocabularies. In particular, our main reference is represented by GEMET, because of its consolidated use within the scientific community and the alignments it already has with other vocabularies.

More specifically, automatic procedures are used to discover and establish exact and sometimes close matches across concepts coming from different vocabularies, while human mediation was necessary to validate the output of automatic procedures and to

\textsuperscript{16} For a description of ontology structure and construction see (Noy and McGuinness 2001; Capuano 2005).
identify both hierarchical and associative mappings when it is not possible to provide a valid equivalence match. This is done to ensure that all concepts included in the ontology have an external reference in at least one vocabulary involved in ontology mapping. With regard to equivalence, using an hybrid method allows to control and manage inexact matches (ex. Homographs, Synonyms), because, as within the same vocabulary, also in cross-mapping, the equivalence relationship can be exact, inexact and partial (ISO 25964-2:2013). The different degrees of equivalence mapping have been expressed by means of the SKOS data model that has been imported in Protégé and its properties skos:exactMatch and skos:closeMatch have been used as AnnotationProperties. For the sake of completeness, some skos:broadMatch, skos:narrowMatch and skos:relatedMatch have been included in our ontology, as can be seen in Figure 5 below.

![Figure 5. Examples of mapping](image)

To overcome differences in vocabulary structures, ISO 25964-2:2013 provides some recommendations concerning the suitable model or combination of models to be used. Indeed, SKOS and the ISO 25964 data model are aligned\(^\text{17}\) so that it is possible to create a mapping compliant to both. As regards our mapping process, we have opted for a combination of the hub model, because our ontology is the core and the external vocabularies act as satellites, and the selective mapping, because links are established only in one direction and solely for concepts used in the ontology.

The vocabularies involved in the mapping partially overlap with the ERA-PLANET thematic strands, as illustrated in Figure 6 below.

![Figure 6. Mapping percentages](image)

The preliminary results, exclusively regarding exact and close matches, therefore show the necessity to extend this mapping involving other vocabularies.

### 4.0 Conclusion

Ontology mapping is a challenge in some strategic domains when a large volume of data need to be managed. The ongoing initiatives and actions highlight the importance of interoperability, discoverability, accessibility and reusability of data. These issues are central to our project because all data must be made available to policy decision-makers, who have the hard task of designing the strategies to be used to support sustainable development. In this sense, the ontology and the matching with external vocabularies,

\(^{17}\) https://www.niso.org/schemas/iso25964.
represent a valid semantic support because they allow to expand and specialize an information request. The ontology evolves continuously, therefore a periodic update is necessary in order to revise mappings and add new links to further vocabularies. To improve semantic interoperability in the KP it is expected to define links to some further existing vocabularies (ex. AGROVOC) and ontologies (ex. Sustainable Development Goals Interface Ontology (SDGIO), Chemical Entities of Biological Interest (ChEBI)) which cover some other areas of interest such as health, agriculture and chemistry, that are strictly related to the environmental domain.

Acknowledgments
This activity was funded by the European Commission in the framework of program “The European network for observing our changing planet (ERA-PLANET)”, Grant Agreement: 689443. The authors would like to thank colleagues from IIA-CNR Florence for their collaboration in defining the ontological model.

References
Fugazza, Cristiano, Sören Dupke, and Lorenzino Vaccari. 2010. “Matching SKOS Thesauri for Spatial Data Infrastructures.” In Metadata and Semantic Research:Research Conference on


A Methodological Proposal Towards Domain Ontology Enrichment

Abstract:
Since the current methods for domain ontology enrichment present some gaps, due to knowledge dynamicity, this investigation aims to develop a methodology for domain ontology enrichment that overcomes the existing methods’ gaps. To address the goal, four steps sustain the research methodology: 1) An exploratory study of Knowledge Organization Systems maintenance and updating; 2) Mapping and analysis of the methods for enriching ontologies, from the literature review; 3) Qualitative content analysis of documents selected in Phases 1 and 2; 4) Development of the methodology for domain ontology enrichment. The result is a novel methodology for domain ontology enrichment, called METHODOE.

1.0 Introduction
Since ontologies are a type of Knowledge Organization Systems and knowledge is dynamic, ontologies must be updated periodically. Unfortunately, most ontologies developers ignore the ontology maintenance and updating area. They focus only on the development of these KOS, ignoring the fact that knowledge can change, terms used to represent concepts become obsolete, new terms emerge, and new scientific discoveries are made. So, ontologies must pursue this knowledge evolution.

An approach to update ontologies is through the ontological enrichment process. The enrichment process aims to expand an already developed ontology with new components (e.g. concepts, relationships, properties, and axioms); in consequence, the domain representation increases its potential. Hence, this study aims to develop a domain-independent methodology for the ontology enrichment process.

The literature presents several proposals for ontology enrichment. These proposals have the following limitations: 1) enrichment only of some ontologies’ components (e.g., Faatz 2001; Faatz and Steinmetz 2002; Valakaros et al. 2004); 2) enrichment based on a very particular data source (e.g., Navigli and Velardi 2006; Amar, Gargouri, and Hamadou 2013); 3) enrichment applied to a specific domain (e.g., Faatz and Steinmetz 2002; Valakaros et al. 2004; Navigli and Velardi 2006; Booshehri et al. 2013); 4) intuitive and non-systematic methods.

The question of this investigative research is: how to develop a methodology for domain ontology enrichment that overcomes the gaps in existing methods?

We assume that literature can provide indications of how we can get the knowledge to develop this kind of methodology. Especially in the literature regarding Knowledge Organization Systems maintenance and in the empirical studies about ontology enrichment.

The motivation behind this theoretical investigation is the concern in improving existing domain ontologies since these instruments are important in the communication, interpretation, and reasoning of knowledge. Besides, ontologies help in the Semantic Web context, favoring the semantic integration between different systems and vocabularies. Likewise, it supports the organization and retrieval of information. Thus, studies are necessary to contribute to the topic of ontology maintenance because ontology researchers still neglect this area.
2.0 The research methodology

The research methodology is sustained in the following four steps: 1) An exploratory study of Knowledge Organization Systems maintenance and updating; 2) Mapping and analysis of the methods for enriching ontologies, from the literature review; 3) Qualitative content analysis of documents selected in steps 1 and 2; 4) Development of the methodology for domain ontology enrichment.

1) An exploratory study of Knowledge Organization Systems maintenance and updating

In this research, we consider the ontology enrichment process as belonging to a more comprehensive area within ontological engineering, which is the ontology maintenance. Thus, we determine the main norms and methods for maintaining Knowledge Organization Systems as a knowledge source, from which we could extract inputs for the development of the methodology we intend to work within this study.

KOSs, such as thesauri, classification systems, and taxonomies, have similarities with ontologies in some aspects. Therefore, some strategies for maintaining and updating these instruments can also be reused in ontologies. So after exploratory research on the topic, we selected the work of these authors: Kim (1973), Soergel (1974), ANSI/NISO Z39.19 (2005), ISO 25964-1 (2011), ISO 25964-2 (2013).

2) Mapping and analysis of the methods for enriching ontologies from the literature review

In this stage, a narrative type literature review was carried out, where the objective was to map the works addressing the ontology enrichment thematic to verify which researches have already addressed this theme. The research was carried out in Information Science and Computer Science databases using the following expressions (in English and Portuguese): Ontology Enrichment, Ontological Enrichment, Ontology Expansion, Ontology Extension, Ontology Specialization, Ontology Refinement, Ontology Enlarge, Ontology Completeness, Ontology Improvement. The search strategy used truncators, boolean operators, advanced search, and specific filters for each database. The period determined in the research was from 1990 to 2018.

After applying the search strategy, duplications and works not related to the scope of this research (by analyzing the title and abstract) were eliminated, obtaining a result of 35 works in total. Then, these 35 works were read using the following exclusion criteria: (1) works dealing with another Knowledge Organization Systems or databases enrichment; (2) works addressing another process and not enrichment (such as ontology learning, evolution); (3) works focusing only in the technique of knowledge acquisition and did not result in the enrichment of ontologies. In the end, 15 studies were considered for analysis in this review, which are: Faatz et al. 2001; Faatz and Steinmetz 2002; Valakaros et al. 2004; Navigli and Velardi 2006; Bendaoud, Toussaint, and Napoli 2008; Carvalho et al. 2010; Barbur, Blaga, and Groza 2011; Petasis et al. 2011; Hashimy and Kulathuramaiyer 2013; Booshehri et al. 2013; Amar, Gargouri, and Hamadou 2013; Booshehri and Luksch 2015; Al-Yahya, Al-Malak, and Aldhubayi 2016; Gómez-Moreno; Mestre-Mestre, 2017; Guerram; Mellal, 2018.

3) Qualitative content analysis of documents selected in steps 1 and 2

Given the recovered and selected documents in the previous steps, 1 and 2, a qualitative analysis of these documents’ content was performed. We selected the main points which deal with maintenance and updating in KOS and based on this, we created categories to organize the extracted information.
The analysis of the documents selected in step 1 demonstrates that when maintaining Knowledge Organization Systems there are a few points to consider: (1) define the person responsible for maintenance; (2) categorize the type of change (which may cover the inclusion of a term, replacement, exclusion); (3) control changes made (like the source of the extracted information, inclusion date); (4) identify the consequences of maintenance for other systems using the KOS; (5) define a periodicity for the maintenance. Therefore, despite being focused primarily on updating thesauri, the way this process works has well-founded information adaptable to the context of ontology enrichment.

The analysis of the 15 selected papers revealed that they employ methods focusing on the technique used for information extraction, the knowledge source, and the type of enrichment performed. Among the identified information extraction techniques, there are statistical analysis, similarity measures, machine learning algorithms, syntactic analysis (part-of-speech, named entity recognition, parsing, stemming, tokenization, lemmatization), formal concept analysis, cluster techniques. Concerning knowledge sources we identified: textual corpus, semantically annotated corpus, thesaurus, ontology, web page content, lexical bases (such as Wordnet), machine-readable dictionary, data in linked data. Regarding the types of enrichment, they highlight lexical enrichment, conceptual enrichment, enrichment of taxonomic relations, enrichment of non-taxonomic relations, enrichment of axioms.

We realize that most works have a narrowed perspective and they approach specific methods for a knowledge domain, consequently presenting the following shortcomings: (1) there is no planning for enrichment; (2) lack of details on how to perform some steps; (3) the methods are empirical, intuitive and not systematic. Notwithstanding these shortcomings, the analysis of these articles and papers still provided valid inputs for the development of the enrichment methodology.

We created an action plan from the inputs generated in steps 1 and 2, which brought up information about what should be considered when maintaining a KOS and also about the main components of the enrichment process (such as knowledge extraction technique, knowledge source, and type of enrichment). This action plan will help in the development of the enrichment methodology, which will appear in the next stage of the research methodology.

4) Development of the methodology for domain ontology enrichment

Information gathered in the previous phases helped to develop an action plan with key strategies for the development of the ontology enrichment methodology. As seen in Table 1, this plan has seven features and 11 strategies that will guide the methodology’s development ensuring it is better managed.

<table>
<thead>
<tr>
<th>#</th>
<th>Functionality</th>
<th>Action / strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Assessment of the need for enrichment</td>
<td>1.1) Develop a step to analyze the objectives of the ontology, users, and competency questions, if any.</td>
</tr>
<tr>
<td>2</td>
<td>Ontology diagnosis</td>
<td>2.1) Describe the possible ways to make the diagnosis.</td>
</tr>
<tr>
<td>3</td>
<td>Knowledge acquisition</td>
<td>3.1) Describe the possible knowledge sources;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.2) Describe the possible techniques for extracting knowledge.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.3) Explain how to extract knowledge.</td>
</tr>
<tr>
<td>4</td>
<td>Knowledge processing</td>
<td>4.1) Explain how to handle the extracted information.</td>
</tr>
</tbody>
</table>
The methodology developed according to the action plan above will be described in the next section, "Results."

3.0 Results
As a result, we present the three phases which unfold in seven steps of the proposed methodology, which we call METHODOE (Methodology for Domain Ontology Enrichment): 1) Pre-enrichment; 2) Enrichment; 3) Post-enrichment. Figure 1 presents a general outline of METHODOE.

This methodology maps the entire enrichment process and makes it structured and organized. This methodology does not intend to describe in detail each of the possible techniques for extracting knowledge to carry out enrichment. Rather, it will be flexible and allow the ontologist to choose the knowledge source and the extraction technique that best fits the domain represented by the ontology since each domain has its particularities.

![Figure 1. METHODOE general outline](image)

3.1 Pre-enrichment
Pre-enrichment is the first phase proposed in METHODOE and has the following steps: (A) Assessment of the need for enrichment; (B) Ontology diagnosis.

In Step A - Assessment of the need for enrichment occurs the analysis to understand if the ontology responds to its objectives, this can be done using various methods from the ontology evaluation area. A very common way to do it is through Competency Questions (CQs). Thus, if the ontology is unable to answer the competency questions elaborated in its development project, there will be indications showing that it needs enrichment. It is also possible to propose new CQs for the ontology. The assessment frequency of the need to enrich the ontology will depend on each domain, or how often
the terminology and the domain evolve. At this stage, a report should be generated, describing the need, or not to enrich the ontology with the necessary justifications.

In Step B -Ontology diagnosis, an examination of the ontology is implemented to find out in which parts the enrichment will be necessary. Again, there are multiple ways to do this. The ontologist can check the entire ontology (each concept and relationship) with a domain specialist guidance and look for possible points where improvements might be necessary. Another possibility is the use of 41 pitfalls for diagnosing ontologies developed by Poveda-Villalón (2016), as they deal with common mistakes made when building ontologies, these errors are indications for carrying out the enrichment activity. Not all pitfalls apply to the enrichment process. In this step, a detailed report (identifying mainly the location of the problem in the ontology structure) should be generated from the diagnosis result.

3.2 Enrichment

After the pre-enrichment phase, the ontology enrichment process is carried out. The steps composing this phase are (A) Knowledge acquisition; (B) Processing of extracted knowledge; (C) Content insertion in the ontology.

Step A - Knowledge acquisition deals with access to the raw material (information) for enrichment, and can also be an iterative step that happens throughout the enrichment process. This stage has three activities: (a) selection of knowledge sources; (b) indication of knowledge extraction techniques; (c) knowledge extraction.

About the Activity a - selection of knowledge sources, there are various knowledge sources, such as domain experts, textual sources (such as articles, books, reports), and other Knowledge Organization Systems (thesauri, taxonomies, glossaries, ontologies). In short, all sources considered qualitatively able to acquire information on the domain represented by the ontology can be recognized as a potential knowledge source. The participation of a domain expert in the survey of these sources can be very useful and should be considered. Moreover, one should consider the diagnostic report done in the Pre-enrichment phase because it can provide inputs to find the knowledge sources. Furthermore, we suggest exploratory research in domain databases and repositories of KOS. After surveying all potential knowledge sources, a table with all sources, and the motive for choosing each should be generated.

Concerning Activity b - an indication of knowledge extraction techniques, it can range from interviews with domain experts, analysis of documents, and KOS. These techniques can be manual, semi-automatic, or automatic using linguistic techniques (Natural Language Processing mainly), statistics, and based on machine learning algorithms. This step must generate a table with the selected techniques accompanied by a motive for choice.

The knowledge source and the technique must have a strong relationship since the chosen source's nature will greatly influence the knowledge extraction technique that will be used. Having chosen the technique, the information is extracted from the selected knowledge sources. In METHODOE, we do not explain which source or technique should be used since this will depend heavily on which domain the enrichment process will be developed. We highlight that one of this methodology's characteristics is to be domain-independent. Thus, one must consider the existing sources in each specific domain.

Activity c - knowledge extraction, refers to the application of knowledge extraction techniques in the selected knowledge sources. Again, the report on the ontology
diagnostic stage is important because it will assist the specific search for knowledge to enrich the ontology. The goal here is not to extract all knowledge regarding the domain represented by the ontology, but to extract the knowledge the ontology has yet to cover. To this end, specific questions to the knowledge sources are made. These questions can be asked through interviews with domain experts, through text analysis (manually, automatically, or semi-automatically) of the domain in pursuit of answers. The manner used to extract knowledge must be associated with the type of knowledge extraction technique and the chosen knowledge source.

Stage B – Processing of extracted knowledge deals with the organization of knowledge acquired in the previous phase and comprises two activities: (a) correlation between diagnosis and extracted knowledge; (b) classification of extracted knowledge.

Activity a - correlation between diagnosis and extracted knowledge tries to relate the gaps identified in Step 1.B (Ontology diagnosis), and the possible solutions found through Activity 2.A.c (Knowledge Extraction). The objective is to facilitate the identification of possible answers to the questions and gaps the ontology presents.

Activity b - classification of extracted knowledge refers to the attempt to group the acquired knowledge into types of enrichment. The types of enrichment are: (1) lexical enrichment, which deals with the acquisition of terminological variations of a concept, synonyms, and definitions in natural language; (2) conceptual enrichment, refers to obtaining new concepts, they can be specific or general; (3) enrichment of taxonomic relations, that deals with the acquisition of 'gender-species' and 'part-of' relationships between the ontologies’ concepts; (4) enrichment of non-taxonomic relations, which deals with all other types of associations that can happen between the ontologies' concepts; (5) enrichment of axioms, discusses the definition of rules to the concepts and relationships, its purpose is the formalization and consequent restriction on the interpretations of the represented knowledge. The results generated by Activity a - correlation between diagnosis and extracted knowledge will be relevant for the classification of knowledge into types of enrichment.

Step C - Content insertion in ontology works with the ontology expansion with the acquired content. This step must be performed by the ontologist or together with the domain specialist. The purpose is to enrich the ontology according to Step B - Processing of the extracted knowledge’s results and correctly insert the content in the ontology. If something remains unclear, it is possible to return to the knowledge acquisition stage. At the end of this step, one must generate a detailed report with all the content inserted in the ontology, the knowledge source, and the inclusion date in the ontology.

3.3 Post-enrichment

The last phase of METHODOE consists of the following steps: (A) Evaluation of the enriched content; (B) Documentation of the methodology.

Step A - Evaluation of the enriched content verifies if the content has been properly inserted in the ontology and if there are any necessary changes to the structure or ontology component since the inserted content can have an impact on its structure. This analysis is performed with the help of the domain specialist. Finally, a document specifying the evaluation result is generated.

Stage B - Methodology documentation refers to the registration of all procedures performed in each stage of the ontology enrichment. This step should happen throughout the entire process, not merely at the end. However, only at the end of the enrichment process, will it be possible to generate the final document of the methodology.
4.0 Conclusions

We presented a novel methodology for domain ontology enrichment based on indications found in the literature related to Knowledge Organization Systems’ maintenance and updating, and in empirical studies regarding domain ontology enrichment. This methodology's fundamental goal is the attempt to organize the entire domain ontologies’ enrichment process. It pursues to do so presenting the pre-enrichment phase steps to be executed before the enrichment itself, and steps to be done afterwards, differing from all methods presented in literature so far. Thus, it contributes so that ontologists can have a systemic and holistic notion of how to improve existing ontologies.

METHODOE is still in its first version, so it needs to be applied in domain ontologies to be diagnosed and validated, which is one of the main limitations of this research. However, we believe the methodology presents a relevant and systematic view on the process of domain ontologies enrichment, surpassing the methods previously developed in some aspects.

As a forthcoming proposal, we intend to validate this methodology in the enrichment of two ontologies from different domains, thus proving the METHODOE's domain-independent characteristic. Furthermore, we intend to develop a practical and exemplified manual for each stage of the methodology.

References


Faatz, Andreas and Ralf Steinmetz. 2002. *Ontology Enrichment with Texts from the WWW.*


Cultural Warrant
Old and New Sights from Knowledge Organization

Abstract:
Culture is a controversial and multi-discursive term in different disciplines. A dimension of Knowledge Organization (KO) appears in the cultural warrant (CW) because KO systems, processes and products show “the values of the cultures involved” (Guimarães et al. 2019). We identified four theoretical elements necessary to understand the CW: the need of adjusting the concept of culture to the KO field; the focus placed on local dimensions of knowledge; the intention to promote biased classifications in favour of minority and relegated social sectors; and the ethical issue expressed in the respect for the integrity of the cultural thought of a community. The suitable techniques for the application of the CW in the processes of construction, evaluation, and revision of knowledge organization systems (KOS) are qualitative. Some methods are common to KO: content, terminological, discourse, and domain analysis techniques. We also propose a categorization of three cultural hospitalities. It is concluded that the concepts of culture and CW are not neutral since they favour forms of knowledge organization that replace the criteria of objectivity and neutrality, by those of cultural pertinence and respect for the values of a community. We also suggest considering the extrapolation of some methodologies from social sciences to study the linguistic behaviour of subcultures in order to improve the CW of KOS.

1.0 Culture, warrant, cultural warrant: consensuses and conflicts
1.1 Culture
Culture is a controversial and notoriously ambiguous term used in different disciplines (anthropology, sociology, politics, feminism, humanities and cultural studies). It is also a “multi-discursive” term, because “it can be mobilized in a number of different discourses” (Hartley 2004, 51). The fact of being multi-discursive explains much of the difficulty to agree on a unique concept of culture. Not surprisingly, more than 150 definitions have been collected in the classic review by Kroeber and Kluckhohn (1952), and all of them are adequate.

Culture is “a term that evolves in the historical period during which it develops,” and therefore, “it has changing contents” (Rodríguez Pastoriza 2006, our translation). Its oldest antecedent is associated with agriculture. In Latin ‘cultus’ means cultivation, cultivated, or treated with care. The semantic background has several positive connotations since crops imply subsistence, regularity, and continuity (Di Tella et al, 2004), the vigorous growth of new forms of life.

The concepts of culture and civilization have been bound together for several centuries, and have been seen from the French tradition of the Enlightenment “as a progressive, cumulative, distinctively human achievement” (Kuper 1999, 5) to which everyone can aspire. Hoggart (1957) and Williams (1958) were the first to propose to overcome the idea of culture as proper of an elite and planted the seed of the so-called cultural studies, which transformed the concept, seen today as “a dynamic concept, always negotiable and in process of endorsement, contestation, and transformation” (Wright 1998, 10).

From an individual perspective, culture can be considered as the measure of a person’s education, manners and knowledge of the world.
But what matters for the objectives and purposes of Knowledge Organization (KO) is culture considered from a social perspective, because it gives an adequate dimension of the information problems and the solutions that have to be addressed, and because each individual (and therefore, each information user) belongs to several cultural communities due to ethnicity, religion, nationality, political thought, habits, and preferences. This was confirmed by Baumann, when he studied five ethnic groups in a London sector, looking for common elements of each culture and found, in addition to those identity traits, what he called “communities within communities as well as cultures across communities” (Baumann 1996, 10). The educational background, professions, and disciplines are also important personal cultural variables.

Even if culture is often defined as a system of shared meanings, Burke (2005) determined that it was difficult to hold this position when large groups, such as nations, were studied. Burke argued that this approach showed the strengths and weaknesses of the Durkheimian model of society, where consensus prevails over conflict (two major topics in culture). He proposed -as an alternative- “the use of the concept of subculture, defined as a partially autonomous culture within a larger totality,” without intending to give an idea of inferiority (Burke 2005, 177).

Burke added that sociologists had dealt with the most visible subcultures (ethnic or religious minorities), those deviant according to the rules of a given society (criminals and heretics), and young people. Historians have also studied groups such as the Jews in Medieval Spain (Pérez 1993) or beggars in Elizabethan London, “but they did not always pay attention to the relationship between the culture of those minorities and those of the surrounding society” (Burke 2005, 178).

Hebdige put an end to this perspective by analysing the process through which dominant mentalities in a society become hegemonic, and treat subcultures in a somehow pejorative way; this gives rise to a tension between the powerful groups and minorities (Hebdige 1979).

Briefly, the following approaches can be taken into account in KO:

i) On the one hand, those that promote consensus values, such as cultural integration and tolerance between cultures. Unesco follows that trend, when it states that culture is conceived “as the set of distinctive spiritual, material, intellectual and emotional features of society or a social group, and [...] it encompasses, in addition to art and literature, lifestyles, ways of living together, value systems, traditions and belief” (Unesco 2001).

ii) on the other hand, those that focus on the analysis of the conflict between different cultural expressions, stating that a dominant culture is usually shielded to strongly resist changes and innovations; thus, they cause different contradictions (higher or elite culture vs. popular culture; hegemonic culture vs. subcultures or minority cultures, accepted cultures vs. rejected cultures, urban culture vs. rural culture, literate culture vs. non-literate cultures).

Besides accepting that classificationists must take into account consensus and conflict between cultures when they develop knowledge organization systems (KOS), it may be more appropriate to give equal importance to the principles and methods that ensure reciprocal tolerance, respect for 'otherness' and the promotion of cultural integration values.
1.2 Warrant

As said in a previous paper, “there are no substantial differences regarding the definition of the notion of warrant” (Barité 2019, 652). In 1986, Beghtol developed such a precise and detailed definition of warrant, that it is still peacefully accepted to this day: “the warrant of a classification system can be thought of as the authority a classificationist invokes first to justify and subsequently to verify decisions about what classes/concepts to include in the system” (Beghtol 1986, 110-111).

More than 30 years after its formulation, it might only be added to that definition that: i) warrants are currently seen as an essential component of three KOS processes: construction, evaluation and revision; ii) we can also consider warrants as tools for terminology selection in contexts other than KOS: information and data systems, big data analytics, social classifications, terminological data banks and specialized dictionaries, among others.

The different existing warrants have recently been compiled and explained (Barité 2018), and a discussion on the usefulness of a general understanding of the problems associated with the characterization and application of warrants (Bullard 2017; Barité 2019) has started.

The fact that in KO literature there is an unequivocal understanding of what a warrant is, its utilities, and applications, is a sound foundation for research.

1.3 Cultural Warrant

While the concept of culture turns out to be multi-discursive and controversial and that of warrant is peacefully accepted in KO, what is left to say about the concept of CW? It has been written that the CW is “a critical activity [that] can be used to evaluate both classification schemes and knowledge fields” (Hjørland and Albrechtsen 1999, 135) because KO systems, processes and products reflect “the values of the cultures involved” (Guimarães et al. 2019, 29, our translation).

It has repeatedly been said that although classifications try to represent the map of disciplines in an objective and neutral way, their schemes are historically and culturally conditioned, since they reflect the social, political and religious thought as well as the state of the scientific evolution of their times and designers’ mentalities (Shera 1961; Lee 1976; Beghtol 1986; González Casanova 1996).

KOS have recurrently been criticized for unequal (Trivelato and Moura 2016), improper (Kua 2004), discriminatory (Furner and Dunbar 2004; Olson 2007) or colonialist (Pacey 1989) treatment, or for making invisible many topics (Olson and Ward 1998) in religion, social sciences, literature or gender studies. In all these cases, the classificationists’ cultural perspectives have been seriously questioned, and this has compelled to look for systematic solutions to these issues.

It was Beghtol who refined and expanded Lee’s original and basic idea (1976), pointing out that CW “posits that every classification system is based on the assumptions and preoccupations of a certain culture, whether the culture is that of a country, or of some smaller or larger social unit (e.g. ethnic group, academic discipline, arts domain, political party, religion and/or language)” (Beghtol 2002a, 45).

For a better understanding, in the following sections, the concept of CW will be broken down into theoretical and methodological aspects.
2.0 Theoretical approach to cultural warrant

In the literature review performed for this paper, we identified four theoretical elements necessary to understand the CW:

i) The relevance of adjusting the concept of culture to the KO field. In the literature of the area, it has seldom been explained on what concept of culture the development of an idea is based on. A broad sense -that may not be acceptable to everyone- is usually taken for granted.

Lee made a relevant contribution when she reviewed definitions of culture from both KO and anthropology literature and compiled the definitions in four families of definitions (Lee 2015).

As culture is a multi-discursive and controversial term, it seems logical to seek an operational definition to be used as a reference for each research in KO, integrating the peculiarities of vocabulary control processes, classification, indexing, tagging and information retrieval on culturally determined topics.

ii) The focus placed on local dimensions of knowledge as opposed to universal approaches. The recognition of the local dimensions of knowledge means addressing the conception of science and technology as a set of ideas, solutions and applications valid in any place and time and alternative thinking that addresses specific beliefs, values, and contexts where knowledge acquires a more immediate legitimacy. Tensions between global culture and local cultures are expressed through language, which always operates as a reference to cultural identity.

Notwithstanding the recognition that KO has traditionally favoured universal systems to ensure some uniformity in the international communication of data on publications, in recent years two insurgent movements against universalistic perspectives have been identified. On the one hand, there is a series of critical studies on the serious limitations that universal systems have for the classification and indexing of issues of local impact. On the other hand, we see the emergence of new KOS - especially thesauri - developed for subject representation of local affairs (Brasil, Ministério de Cultura 2006).

Some disciplines are supported by the cultural dimensions of a field (law, history, cultural anthropology, music, literature, and arts), or by approaches built by various disciplinary cultures; they were the first to raise issues of subject representation. For example, the “mate” culture -widespread in Paraguay, Argentina, Uruguay, and southern Brazil-, has its own body of literature and terminology. Its status requires the creation of classification systems or mini-thesauri to adequately represent the specific documentation on the “mate” infusion and the cultural elements (rituals, objects, procedures, codes of conduct) that surround it.

Likewise, different cultures classify birds differently, and this is studied in ethnobiology (Berlin 1992) and in biological systematics, where different paradigms such as numerical taxonomy and cladistics are competing perspectives.

iii) The intention to promote biased classifications in favour of minority and relegated social sectors. This approach implies accepting the distinction between dominant and subjugated cultures that coexist in the same society. Although this may be an excessive simplification, part of the social conflicts may be related to the non-peaceful coexistence between different cultural communities.

Minority cultures strongly protect their cultural characteristics (language, customs, and ways of understanding reality). They result in followers, theories, documentation,
regulations, objects, liturgies and procedures that require visibility in KOS. Members of minority cultures often develop strong religious, ethnic, ideological and/or philosophical cohesion, especially when the dominant or hegemonic culture reacts showing indifference or real discrimination.

The DDC and UDC classification systems have validated, in many cases, anachronistic or ideologically biased, tendentious, and even offensive constructions of social and cultural knowledge areas. The CW intends to compensate for these imbalances, and maintain the integrity and identity of minor or ‘minoritized’ cultures (Olson 2007).

iv) The ethical issue expressed in the respect for the integrity of the cultural thought of a community. By involving in subject representation the visibility of particular groups, showing their patterns of coexistence and communication, the CW introduces the ethical factor in KO (Beghtol 2002b; Guimarães et al. 2008). This is especially important when attention is focused on the demands of social movements or people who promote ideas based on new ethical presuppositions, or the deconstruction of hegemonic cultural perspectives or forms (Garcia Gutiérrez 2007).

The identification of the CW with essential ethical precepts regarding the recording, availability, access to and retrieval of information in the most open and free way allows us to place it in the pragmatic epistemological approaches to KO advocated by Hjørland (Hjørland 1999, 2013; Barité 2019).

In this context, the concern for not enhancing the hegemonic aspects of a society and considering alternative interpretations of reality forms should be central theoretical orientations in the application of the CW.

This will imply the use of non-discriminatory or non-inclusive indexing terms and politically correct expressions that take into account all the possible perspectives for the formulation of a culturally determined topic or issue. In particular, terminology selection should be endorsed by movements or social groups considered as potential users of an information service or system.

The different dimensions of culture (culturality, culturalism, multiculturalism, interculturality, and transculturality) - studied by Boccato and Biscalchin (2014) - will have to be taken into account among other aspects for the development of the CW concept. The customized and comparative treatment of these dimensions results in a better disaggregation of the concept of culture, but methodological precautions have to be considered.

3.0 Methodological approach to cultural warrant

From a methodological perspective, the suitable techniques for the application of the CW are qualitative. Some methods are common to KO:

Content analysis: a classical tool used for indexing and tagging (Krippendorff 2004; Hsieh and Shannon 2005) which allows identifying culturally marked expressions to be used as appropriate indexing terms for specific user communities.

Terminological analysis: used to create KOS biased towards certain aspects of social and human sciences, or the understanding and judgment of situations where the parameters of two different cultures may be in conflict. The paper by Benyaich (2014) on the consequences of the relative incompatibilities between family law in Spain and Morocco is a good example.
Critical discourse analysis: the assumptions of this methodology listed by Kress, one of the pioneers, have several points of support for the CW. Kress assumed that language is a social phenomenon. Therefore, individuals, institutions and social groups convey meanings and values through language in an organized way. In this context, texts are fundamental units of language in communication. Readers are not passive recipients of texts; on the contrary, they interact in various ways with those texts based on their cognitive structures, perceptions, and cultural patterns (Kress 1989).

Van Dijk (1999) developed categories and procedures for the critical study of the discursive reproduction of domination and hegemonic thinking in societies. These categories and procedures can also be articulated with the CW methods.

Domain analysis: Under this generic name, Hjørland gathered eleven techniques for mapping areas of knowledge from different perspectives and bases of analysis (Hjørland 2002). Some domain analysis techniques -whether qualitative or quantitative- can also be fruitful for the CW, for example: the creation of literature guides; the construction of special classifications and thesauri; empirical user studies; historical studies; epistemological and critical studies; database semantics and discourse studies; and even bibliometric studies.

The CW can be used throughout a KOS specialized in cultural issues, such as the Art & Architecture Thesaurus (Getty Research Institute 2017) or the Brazilian Folklore and Popular Culture Thesaurus (Brasil, Ministério de Cultura 2006).

Some culturally oriented terms can also be inserted into a pre-existing scheme, to make visible a certain perspective. This implies the need to intervene in that pre-existing system, adding specifications in precise areas of the schemes, or inserting alternative tables of local value.

To that end, it is necessary to make changes in the pre-existing system, adding specifications to certain areas or inserting alternative tables of local value.

Beghtol (2002a, 2002b) introduced the principle of hospitality that promotes and values the insertion in the KOS of new, alternative or local specifications. The 'cultural hospitality' implies the capacity to create procedures so that KOS are permeable to different cultural perceptions and conceptions.

In one of her papers, Beghtol pointed out that the cultural hospitality “needs to be debated, assessed and tested further to assess its potential for effective implementation” (Beghtol, 2002a p. 48). In response to that call, we propose here the initial categorization of three forms of cultural hospitality:

i) The technique of creation of paradoxical spaces (Rose 1993; Olson and Ward 1998) to insert gender terminology into the Dewey Decimal Classification (DDC) tables. The method could also be used in any KOS to insert other culturally determined terms (by race, religion or subculture).

This technique proposes to subdivide a general concept in order to incorporate particular concepts that have been omitted. As an example, Olson and Ward (1998) subdivided 'Economic basis of labor' to include the topic 'Unpaid labor', as follows:

331.116 Economic basis of labor (recorded in DDC)
331.116 2 Unpaid labor (incorporated by the creation of a paradoxical space)
331.116 3 Paid labor (recorded in DDC)
ii) The creation of schemes biased towards certain cultural orientations within pre-existing or autonomously built classifications. In a recent paper about bias in KO, Colombo and Barité (2015) identified and categorized three forms of bias in KOS: positive bias, negative bias and neutral bias. Of the three, the positive bias, understood as the premeditated will to orient the terminology of a KOS in a certain direction (Buddhist thought, Marxist conception, an evangelical approach, a feminist stance in the choice of terms) is a door of access to a cultural perspective with its own identity.

iii) The local adaptations of universal classification schemes to represent the specific characteristics of a country or region. Local adaptations of universal classification systems are a traditional way of expressing cultural hospitality, insofar as they seek to address the peculiarities of a country or region, its geography, administrative division, literature or history. These adaptations can be designed by those officially responsible for the KOS (Beall 2003; Choi 2018), or they can be of local nature, or internally generated in an information system, or a library system or network, with the advantages and disadvantages it entails.

4.0 Conclusions

The term 'culture' has been used in many different contexts, by various disciplines, with different scopes, thus becoming an essentially ambiguous, controversial and changing concept. Neither anthropology - its field of origin - nor the development of cultural studies have been able to overcome this ambiguity, or reach definitive agreements on its meaning, since each discipline loads the concept with original contents, according to its specific needs. The notion of warrant, instead, has a wide and peaceful consensus in KO.

It can be said that the CW is at the midpoint: while some specific theoretical, methodological and application guidelines have been drawn up, they have been established in a somewhat dispersed and sporadic way in specialized literature. Besides, the reasoned extrapolation of concepts related to culture and cultural aspects from the disciplines that have most discussed these issues to KO still seems insufficient. A more comprehensive work, more case studies, and the promotion of theses that include KO in an orderly way, as well as methodological aspects of anthropology, cultural studies, history, and sociology are required.

It is concluded that the concepts of culture and CW are not neutral. They favour knowledge organization forms that replace the objectivity and neutrality criteria with those of cultural pertinence and respect for community values.

It can be stated that the CW is more directly related to pragmatic epistemological approaches to KO, and, therefore, the ethical factor is naturally included in the CW. The need for KO to have its definition (or definitions) of culture is reaffirmed. It is proposed to use the subculture concept as an operational approach because the CW can possibly be more useful to promote the terminology and expressions of minority or relegated cultures.

In other situations, a broader conceptualization of culture may be necessary, especially for the subject representation of the elements of two or more cultures coexisting in the same territory where it is necessary to find linguistic mechanisms for social, political or religious integration.

The best methodological solutions will surely arise in each particular case. Perhaps it
will be possible to extrapolate some methodologies used to study the linguistic behaviour of subcultures. It definitely seems that qualitative methods are the most suitable tools to work with the CW. In this paper, it has been possible to systematize a set of useful methodologies for the CW and to propose a first categorization with three forms of cultural hospitality.

The CW contributes to reaffirm the identity of local cultures and neutralize the acculturation effects associated with globalization and political-economic processes of social exclusion from KOS.

It is essential to privilege the analysis of users’ relationship with indexing terms and documents organization involving their cultural baggage, their way of understanding reality and assimilating established knowledge.

It is important to point out the integrative and democratic role that the CW can give to our area of knowledge, as it postulates tolerance between different cultures and respect for the cultural integrity of subcultures inserted in our societies.

New questions arise about the CW and its purposes, and they require new answers: should it be a tool to prevent or correct deviations from the culturally appropriate, acceptable or tolerable issues? Should we be satisfied if we obtain politically correct tags? Or should we be involved with the cultural, political or ideological conceptions of social movements that claim a different reality in terms of language? The emphasis on consensus (appeal to cultural integration) or conflict (preferential attention to subjugated cultures or social movements) will give the answer in each case.

Acknowledgement
The authors thank the reviewers for their important suggestions to improve this paper.

References


Maria Teresa Biagetti – Sapienza Rome University, Italy

Bibliographical Relationships in Knowledge Organization Systems
A Historical-Theoretical Perspective

Abstract:
The study endeavours to analyse the bibliographical relationships provided by some prominent cataloguing codes in the 19th century that is, the codes by Panizzi, Jewett and Cutter, with the aim of highlighting the deep organization of cross-references. The bibliographical relationships in ancient catalogues are compared to the bibliographical relationships provided by the FRBR model and analysed in relation to the functionalities offered in linked open data format built by libraries. The study remarks that in the exposition in LOD format of catalogographic data by Libris, data.bnf.fr and datos.bne.es, all the relationships provided by ancient catalogues are not provided.

1.0 Introduction
Bibliographic relationships are one of the central issues of Knowledge Organization and of cataloguing theory. Derivative relationship brings into being different editions and translations of a work and a further class of derivative textual works, such as amplifications, extractions, commentaries, adaptations, and creates families of works from a common progenitor. The frequency and distribution of derivative works has been analysed by Smiraglia and Lazer (1999). In addition to the descriptive, sequential, whole-part and accompanying relationships, Tillet too drew attention to derivative relationships, such as different versions of a work, editions, revisions, summaries, adaptations, new works based on the earlier work, changes of genre, as dramatizations (Tillet 1989; 2001; Green 2001). Also in the new models for organizing bibliographic data, such as FRBR (IFLA 1998) and LRM (IFLA 2017), bibliographic relationships are crucial. The analysis by Noruzi (2012) reveals that there is an important, albeit not complete, congruence between the categories defined by FRBR and Tillett’s categorization. Additionally, Arsenault and Noruzi (2012) focused on Work-to-Work bibliographic relationships among Canadian publications, and highlighted the frequency and percentage of Work-to-Work bibliographic relationships in some categories, such as supplement, successor, transformation, and adaptation.

RDA <https://www.loc.gov/aba/rda/> (American Library Association 2010) provides guidelines to set up bibliographic relationships between related Works, considering the derivative, the whole/part, and the sequential relations; moreover, it offers lists of relationship designations to be used. In building data set in LOD format, bibliographic ontologies or data models are used, such as Bibo (The Bibliographic Ontology) <http://bibliontology.com/> and Bibframe <http://bibframe.org> (The Library of Congress 2016), in which bibliographic relationships are essential, and expressed in the form of object properties.

2.0 Methodology
The main issue of this study is to pinpoint more exactly bibliographical relationships in bibliographic entities and to highlight the importance of thoroughly analysing the structure of catalogues of prominent libraries in the past centuries, because they provide
a valuable source to study bibliographical categories and relationships involved also in
the realization of modern tools for retrieval. The aim is to provide a more thorough
model of bibliographic relationships to use in particular in LOD built by libraries, with
the purpose of emphasizing the relationships not offered in the modern models. There-
fore, an analysis of a sample of portals of data sets in LOD format is offered with the
aim of analysing the relationships used and provided for searching.

A historical-theoretical survey about the most significant catalographic tools in the
19th century is presented, with the purpose of verifying the structures used to underline
the bibliographic relationships.

3.0 Bibliographic relationships in the catalogues of the 19th century

Some catalogues in the 19th century provide cross-references to connect Works and
Expressions, Works and Authors, Authors and Texts, Author and Author. In this paper,
I consider only a few of them. In The British Museum’s *Rules for the compilation of the
catalogue* (1839), issued within the first volume of the *Catalogue of printed books in
the British Museum* (1941) and for the most part a great deal of effort by Antonio
Panizzi, the Rules LXI-LXII provide a set of cross-references (some of them already
offered by the British Museum catalogue 1787 and 1813-1819) that allow to offer links
between Authors and Works: editors, co-authors and authors of continuations, transla-
tions, comments and biographies. An in-depth analysis of the ancient cataloguing rules
is presented by Biagetti (2001).

In the Report of the 21th February 1839 addressed to the Trustees of the British Mu-
seum (Appendix n. 10 1850) which provides the synthesis of the catalographic regula-
tions established by Panizzi himself, is presented the network of references that should
be realized to link Works, Texts and Authors and to connect an author’s work to all the
authors of the different texts and of illustrations, or to the editors and commentators or
the person who is continuing a work. Table 1 shows a synthesis of the Rules LXI-LXII
for cross-references, some examples from the Appendix n. 10 (1850), p. 192 and my
explanation between brackets:

| The editor’s name and the work edited. | “Garret (William). See Floddon field. The battle, etc. 1822, 8°” (Garret is the editor of the edition of the work about the battle of Flodden field). |
| The author’s name of a biography issued within an edition of a work and the work’s title. | “Campegius (Symphorianus). Arnaldi vita. See Arnaldus or Arnoldus de Villanova. Opera etc. 1520. fol” (Symphorien Champier issued Arnaldi Vita within the edition 1520 of the Works by Arnaldo da Villanova). |
| The co-author’s names and the work’s author and title. | “Fletcher (John). See Beaumont (F.). Comedies and tragedies. 1674. fol” (Fletcher and Beaumont are co-authors). |
| The author’s name of a continuation of a work and the work’s author and title. | “Gaertner (Carolus Fridericus). Supplementum carpologiae. See Gaertner (J.). De fructibus et siminisibus. 1788-1805. 4°” (Gaertner C. F. is the author of the supplement published in 1805 to the Joseph Gaertner’s work). |
| The name of a translator of a work and the work’s author and title. | “Moir (George). See Schiller (F) Wallenstein. 1827. 8°” (George Moir is the translator in English of the Schiller’s work). |
| The name of a commentator of a work and the work’s author and title. | “Sullivan (Arabella). See Ogle (B.) Lady Dacre, Recollections…1833. 8°” (*Recollections of a chaperon* is a histories’ collection known as by Barbarina Ogle, Wilmot, Brand (Lady... |
Dacre, English poet and Arabella Sullivan’s mother, who in this case is considered to be the commentator of the work, but probably is the real author).

It is important to notice that also semantic links are provided: the Rule LXIII connects the name of a person who is the subject of the biography and the name of the biographer; the Rule LXV links the name of an author whose work has been analysed within a work of another author and the name of the latter. Table 2 shows a synthetic explanation and examples from the Appendix n. 10, with my comment.

<table>
<thead>
<tr>
<th>Table 2 Relationships and examples from the Panizzi’s Rules for the British Museum</th>
</tr>
</thead>
<tbody>
<tr>
<td>The person subject of a biography and the biographer.</td>
</tr>
<tr>
<td>“Rousseau (Jean Jacques). Vie. See Barruel de Beauvert (A. J.) Vie de J. J. Rousseau etc. 1789. 8°” (J. J: Rousseau is the subject of the biography by Barruel de Beauvert).</td>
</tr>
<tr>
<td>The author whose work has been analysed into a work of another author and the author’s name.</td>
</tr>
<tr>
<td>“Martialis (Marcus Valerius). See Calderinus (D.) Commentarii in M. 1474. 4°” (Calderinus is the author of the comment).</td>
</tr>
</tbody>
</table>

Earlier, Antonio Panizzi offered a similar structure of cross-references in the classed catalogue of scientific works of the Royal Society\(^1\) published in 1836. In this case, references are used to link editors and commentators to the works, and discovered authors of anonymous works to the works. However, the most relevant thing is to use cross-references with the aim of indexing works published within other publications or within miscellaneous works, reports or comments issued together with the publication to which they are linked, additions and supplements. Moreover, as this is a classed catalogue, the works concerning more than one discipline are linked to each relevant class by cross-references. Table 3 shows a synthetic explanation, examples from the Royal Society Catalogue, 1836, and my comment.

<table>
<thead>
<tr>
<th>Table 3 Relationships and examples from the Panizzi’s Royal Society Catalogue</th>
</tr>
</thead>
<tbody>
<tr>
<td>The author of a work inserted within Another work and the work.</td>
</tr>
<tr>
<td>“Cusa or Cusanus (Nicolaus de). De quadratura circuli deque recti ac curvi commensuratione. See Regiomonte (J. M. de). De triangulis omnimodis. Fol. 1533” (Cusano issued his work as an addition to the Regiomontano’ work).</td>
</tr>
<tr>
<td>The author of a report related to a work and the work.</td>
</tr>
<tr>
<td>“Ampère (A. M.) Rapport sur une mémoire de M. Bérard. See Bérard (J. B.). Méthodes nouvelles pour déterminer les Racines etc. 4° 1818”. (Ampère is the author of the report).</td>
</tr>
</tbody>
</table>

Charles Coffin Jewett (1852) published *On the construction of catalogues of libraries, and their publication by means of separate, stereotyped titles*. Jewett devised a system to use stereotype plates to print catalogues of American libraries with uniform headings, under the direction of the Smithsonian Institution. To this aim, Jewett prepared the code for cataloguing to be used by the Smithsonian’s library, based on the Rules of the British Museum. The code provides cross-references for all the authorial responsibilities in a work: translators, commentators, editors, continuators, also from the name of any author whose work is contained in a collection. Semantic references are provided linking

---


© Ergon - ein Verlag in der Nomos Verlagsgesellschaft
the name of a person subject of a biography or whose work is subject of a commentary (also without the text), to the work. Table 4 shows some explanations and examples from the work by Jewett (1852).

Table 4 Relationships and examples from the Jewett’s work

<table>
<thead>
<tr>
<th>Type of Relationship</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>The author’s name of a biography issued within an edition of a work and the work’s title.</td>
<td>“Bentick (Lord George). See Disraeli (Benj.). Biography of Lord Geo. Bentinck.”</td>
</tr>
<tr>
<td>The co-author’s names and the work’s author and title</td>
<td>“Sievrac (Jean Henri). See Cobbett (Wm.). Roman hist. in French and English; the Fr. by J. H. Sievrac.”</td>
</tr>
<tr>
<td></td>
<td>(The reference is to “Cobbett (William). Elements of the Roman history, in English and French, from the foundation of Rome to the battle of Actium; selected from the best authors, ancient and modern, with a series of questions ... The English by William Cobbett; the French by J. H. Sievrac. London, 1828. 12º”)</td>
</tr>
<tr>
<td>The author’s name of a continuation of a work and the work’s author and title.</td>
<td>“Marleborough (Henry). Ancient Irish histories.—The chronicle of Ireland. By Henry Marlebvrrogh; continued from the collection of Doctor Meredith Hanmer, in the yeare 1571. Dublin, 1809. 8º”.</td>
</tr>
<tr>
<td>The name of a translator of a work and the work’s author and title.</td>
<td>“Taylor (William). See Oriental his. mss. in the Tamil language; transl. with annotations by Wm. Taylor.”</td>
</tr>
<tr>
<td></td>
<td>(The reference is to “Oriental historical manuscripts, in the tamil language: translated; with annotations. By William Taylor, missionary. [...] Madras, 1835”).</td>
</tr>
<tr>
<td>The person subject of a biography and the biographer.</td>
<td>“Alexander, the Great. See Curtius Rufus (Quintus). De rebus gestis Alexandri Magni.”</td>
</tr>
</tbody>
</table>

In the *Rules for a printed dictionary catalogue* (1876), based on the catalogographic practice followed setting up the first volume of the Boston Athenaeum Library, Charles Ammi Cutter established cross-references for all the authorships of a work, such as editors, translators, including designers, painters, cartographers and, in particular cases, engravers; moreover, for commentaries, continuation and indexes of a work. The Rules by Cutter are well known and it is not essential a complete explanation of the categories adopted. Moreover, they are in part similar to those by Panizzi.
4.0 FRBR, LRM, EDM and bibliographic relationships

FRBR provides bibliographic relationships between Entities of the first group: Works, Expressions, Manifestations and Items in order to increase information and help finding linked Entities. Logic relationships are used to connect the second group, Entities, Persons and Corporate bodies, to the Entities of the first group. Semantic relationships are provided to link Works that concern the same topic, and a work of critic of a literary work. All the Entities of the three groups may be subject of a Work.

Considering Work-to-Work relations, the most significant relationships in FRBR are sequel, supplement, concordance, summarization, transformation, imitation. Relationships between Expressions of the same work are Abridgement, Revision, Translation, Arrangement (music); between Expressions of different Works, there are relationships such as Successor, Supplement, Complement, Summarization, Adaptation, Transformation, and Imitation.

In the FRBR, “relationships are examined in the context of the entities defined for the model, i.e., they are analysed specifically as relationships that operate between one work and another, between one expression and another, between a manifestation and an item, etc.” (IFLA 2009, 55)

It is important to bear in mind that, however, the network of relationships concerns mainly transformations of a Work. Considering the attributes, the statement of responsibilities is highlighted only at the Manifestation level: authors, translators, editors, compilers. Knowing an attribute of responsibility related to an Expression of a Work, for instance the editor of a particular edition of a Work, it is not possible to find the Work, running through the web of relationships.

In the LRM (2017, 63) the relationships of second level, in particular for Works, mainly include relations of subject, whole-part (component part), priority (logical continuation), complement (or companion), inspiration for (source of ideas), transformation (change of literary form); considering Expressions, are included relations of whole-part, derivation, aggregation; and in case of Manifestations, relations of whole-part and reproduction.

In the EDM, Europeana Data Model for Europeana Collections (2017), the bibliographic relationships are presented in the form of properties, including “Is Derivative Of”, “Is Similar To”, “Is Next in Sequence To”, “Is Related To”, “Is Representation Of”, “Is Successor Of”, “Contributor”, “Creator”, “Is Replaced By”.

5.0 Data set in LOD format

In this work I considered a little sample of data set provided by the Swedish Libris, the Biblioteca Nacional de España (Datos.bne.es), the Bibliothèque Nationale de France (data.bnf.fr). A mapping between the model proposed by the rules for cross-references provided by British Museum, by Jewett and by Cutter, and the realization of data set in LOD format by some influential libraries, is presented. The aim is to emphasize that LOD format realization by libraries could offer a more detailed and a deeper organization of information using a larger amount of bibliographic relationships, following the ancient catalogues’ model.

Libris <http://libris.kb.se/>

The functionalities for searching in the Swedish OPAC based on LOD technologies offer the possibility to find all the works of an author – for instance, August Strindberg – including monographs and some article about him. Searching for a work – for instance,
Fröken Julie – the user can find all editions in original language and translations assembled by language.

Datos.bne.es, (the beta version) <http://datos.bne.es/inicio.html>:

The functionalities in the Spanish portal allow to find a work – for instance, Don Quijote de la Mancha – and all the different editions also translated in other languages; moreover, they allow to find works about the work considered. Searching for an author – for instance, Miguel de Cervantes Saavedra – the functionalities permit to find the entire author’s works, the works about him, including some biographies and, in case, the works attributed to the author.

Data.bnf.fr <https://data.bnf.fr/>:

The portal offers a great number of functionalities based on semantic web technologies. Searching for an author – for instance, Victor Hugo – the user finds all the works in order of decreasing dates, the musical and iconographic works, manuscripts and archive documents, theatric works of which he is the author of the text; moreover, the works about Hugo are listed, split in categories, such as video, films, images, archive documents; finally, works and persons linked to Hugo are listed: co-authors, designers, engravers, librettists. Selecting a textual work – for instance, Les Misérables – all the editions in French are presented; moreover, the films, the registrations, the theatrical performances based on the work, monographs about the work, documentaries, and virtual shows. Furthermore, data.bnf.fr allows to find a great number of authors linked to the work: editors and commentators, translators, actors, scenographers, producers, and many others.

6.0 Discussion

From the analysis, it appears that in Libris, Datos.bne.es and Data.bnf.fr derivative relationships, such as the categories concerning supplement, continuation and abridgements of a work are not thoroughly provided. On the contrary, these relationships were suggested – albeit restricted to continuation of a work – by Panizzi and followed by Jewett, and have been considered by FRBR. Moreover, there is the lack of the possibilities offered by the set of cross-references provided by the prominent catalogues of the 19th century. For instance, the possibility to connect the person who is the subject of a biography and the name of the biographer (except for Datos.bne) or to link works or reports inserted within another work to the host work, as was suggested by Panizzi’s Catalogue of the Royal Society.

Data.bnf.fr provides the major number of connections based on the FRBR model and proposes the most thorough set of bibliographic information. However, it does not stress on works published within other publications or in miscellaneous works, as suggested by Panizzi. Table 5 shows a mapping between, on one side, a selection of relationships offered by the ancient cataloguing codes of the 19th century and, on the other side, modern data models, such as FRBR/LRM and data exposition in LOD format provided by three prominent libraries.
Table 5 Mapping between modern data models/LODs and ancient catalogues’ relationships

<table>
<thead>
<tr>
<th>Relationship Description</th>
<th>FRBR-LRM</th>
<th>LIBRIS</th>
<th>DATOS. BNE</th>
<th>DATA. BNF</th>
<th>EDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>*The Editor / the work edited</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*The Author of biography issued within a work / the work’s title</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*Co-authors / work’s title</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*The Author of a continuation of a work / the work</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*The Author of a translation of a work / the work</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*The Author of a comment of a work / the work</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*The subject of a biography / the biographer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*The author whose work has been analyzed into a work / the work</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>**The author of a work inserted within another work / the work</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>**The author of a report related to a work / the work</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>***The author of a work published in a collection / the collection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Panizzi-Rules for British Museum, Jewett’s code and, in part, the Cutter’s Rules.
** Panizzi’s Royal Society Catalogue.
*** Jewett’s code.

7.0 Conclusion

The abundance of bibliographic relationships provided by the ancient catalogues through the skillful use of cross-references and links among responsibilities and Works or Expressions, persuades us that it is essential to analyse more thoroughly the structure of important ancient catalogues. In some cases, ancient catalogues show an advanced model of bibliographical relationships, and some connections they allow are not provided either by the FRBR model. This is the case of the Panizzi’s rule for the catalogue of British Museum regarding the link between the name of authors whose work is commented by another author and the name of the latter, or of the rule concerning the use of cross-references to index works published within other publications or within miscellaneous works. The prominent ancient catalogues are a relevant source to study bibliographic relationships involved also in setting up modern tools for searching resources, such as the tools provided by linked open data technologies. Semantic technologies and also bibliographic ontologies should allow to highlight a major number of connections between bibliographic entities. In particular, the works inserted within other works, the reports published within another work and the works issued in collections should receive a major opportunity to be highlighted.

References


Appointed to Inquire into the Constitution and Government of British Museum; with Minutes of Evidences, [...] London, William Clowes and sons.


Europeana. 2017. *Definition of the Europeana Data Model v5.2.8*. 


https://tinyurl.com/y98na59f


https://www.loc.gov/bibframe/


Integrative Levels Classification as a Networked KOS
A SKOS Representation of ILC2

Abstract:
Recently, there is a need to move knowledge organization systems (KOS) to online applications, by using Semantic Web technologies, in order to optimize indexing and searching. The present paper reports the representation of the Integrative Levels Classification (ILC) as a networked KOS, through conversion of its second edition into the W3C standard SKOS (Simple Knowledge Organization System) format.

1.0 Introduction
In recent years, there is an increased need to move knowledge organization systems (KOS) to online applications, such as library catalogues or research data repositories, to optimize indexing and searching. Such need leads to represent traditional organization of concepts in the syntax of Semantic Web technologies (Binding and Tudhope 2016; Trzmielewski and Gnoli 2019). As Peponakis et al. (2019) highlight, enumerative disciplinary classifications and subject headings are harder to represent as machine processable and expressive semantic networks while thesauri are more suitable for this purpose.

Therefore, it is also interesting to observe which solutions may be adopted with freely faceted classifications, which have a richer structure closer to that of thesauri and a greater expressive power. The present paper reports on the representation of the Integrative Levels Classification (ILC) as a networked KOS, through conversion of its second edition into the W3C standard SKOS (Simple Knowledge Organization System) format.

The SKOS work on standards for thesauri and other knowledge organization systems grew out of the EC FP5 SWAD-Europe project. The aim was to facilitate the migration of KOSs to the Semantic Web, and work was carried forward by the W3C Semantic Web Best Practices and Deployment Working Group. The SKOS standard is published as a W3C Recommendation (W3C 2009). While SKOS was designed with thesauri primarily in mind, the availability of a relatively simple and accessible standard, expressible in RDF, has undoubtedly contributed to a major interest in KOSs generally for Semantic Web and linked data application development and also the mapping (and linking) of one KOS to another. Representation in SKOS (and RDF) exposes KOSs to a wide potential audience of developers and users. This was the rationale for investigating the representation in SKOS and making available a machine readable version of ILC.

2.0 The features of ILC
The Integrative Levels Classification is a general KOS that is being developed since 2004 by an international team of scholars, including the authors of this paper. It draws from the tradition of faceted bibliographic classifications as developed by Ranganathan...
and the Classification Research Group. However, it differs from these mainly for listing phenomena — such as iron, lakes, trade unions or orchestras — instead of disciplines — such as chemistry, geography, economics or musicology (Gnoli 2016).

This has important consequences, both theoretical and applied. One of them is that the same classes can be applied to bibliographic records (e.g. an article on bagpipes), museum objects (e.g. a bagpipe specimen), products (a bagpipe model offered in a maker website) and so on, possibly combined with additional dimensions (“bagpipes, in articles”; “bagpipes, in museums”;...: see Gnoli, Park, and Ledl, 2019).

Another original feature is that ILC facets are not only special facets limited to a specific main class (e.g. the processes of biology, or the materials of mining), but also free facets that can be used to connect any two classes from the whole spectrum of knowledge (e.g. cervid populations affected by road traffic). This KOS variety, described by Austin (1976) as freely faceted classification, offers a powerful expressivity very similar to that of a full language; at the same time, it implies a certain amount of syntactic complexity that is more demanding to be represented carefully as linked data (see next section).

The first stable edition of the system (ILC1) was published in 2011 and consisted of 7,052 classes and facets. In September 2019, the developing new edition has been frozen to become the second stable edition (ILC2), consisting of 10,845 classes and facets (Gnoli 2020).

Compared to ILC1, it has evolved in some renamed or moved main classes, better development of many subclasses, rearrangement and new definitions of various facet categories, distinction between facets by nature (“wheels” as parts of vehicles) and facets by function (“vehicles, with wheels”), and other details in notation. These changes are described by Park et al. (2020). Specific fields for mapping between different ILC editions, and between ILC and the Dewey Decimal Classification, are provided in the ILC MySQL database.

ILC features involve a rich semantic structure with many components: basic classes (a-y), common facets (0-9), special facets (90-99), expected foci, deictics (A-Z), etc. While not all these structural components are provided for in the standard SKOS format, good compromises and solutions can be found for many of them (Gnoli et al. 2011).

3.0 Procedures

It was necessary to transform the working representation of ILC2 used by the editorial team to SKOS and RDF. We were able to draw on previous experience by the Hypermedia Research Group at the University of South Wales with publishing national UK heritage thesauri (Heritage Data n.d.) as SKOS based linked data (Binding and Tudhope 2016).

In order to generate a SKOS representation of ILC2, it was necessary to transform the relational (MySQL as exported into CSV) expression of the ILC2 classification system. This was achieved using the STELETO transformation tool developed previously (Binding, Tudhope and Vlachidis 2018). STELETO converts input data to any textual output format via a user-defined textual template. It is a cross-platform command line application (open source) that performs bulk transformation of delimited text tabular data into other textual formats via a custom template (Binding 2019).
Due to the complexity of a faceted classification system such as ILC, bespoke rules were added to the process for ILC purposes. For example, it was necessary to derive the hierarchical structure of the classification from the notational codes. Database fields for synonyms and descriptions of classes, of facet indicators and of foci have been treated variously in order to obtain meaningful labels. Mappings to DDC classes are available for all ILC main classes and for most 3-digit subdivisions of DDC (000-999). These have been linked to OCLC DDC URIs.

The following solutions have been adopted:

- records having purely alphabetic notation values (basic classes) are modelled as skos:Concept
- records having purely numeric notation values (common facets) are modelled as rdf:Property, using the notation to determine the subproperty/superproperty relationships. Single number notations (i.e. the fundamental categories) are sub-property of skos:related. These properties are modelled with domain and range specified as skos:Concept.
- records having a combination of alphabetic and numeric notation (special facets) are also modelled as rdf:Property with the domain being the alphabetic part of the notation and the range being the value from the ‘foci’ field (if present, otherwise skos:Concept). E.g. for \texttt{m981} (“aged organisms”) domain is \texttt{m} (“organs”) and range is \texttt{an} (“quantities”), super property is then \texttt{m98} (“developmental stage”).

4.0 A metaphysical question: what is the top class of all phenomena?

A basic SKOS relationship is skos:broader, by which any class can be related to its parent class. For example, \texttt{wi “pots”} has a skos:broader relationship to \texttt{w “artifacts”} — and vice versa, \texttt{w “artifacts”} has a skos:narrower relationship to \texttt{wi “pots”} and other subclasses. We generated these relationships in automatic ways by exploiting the expressivity of ILC positional notation, where every additional digit means an additional rank of specificity.

Once we came to the main classes expressed by a single letter, such as \texttt{w “artifacts”} or \texttt{h “celestial bodies”}, we had to decide whether these in turn have any skos:broader relationship. ILC2 also has a class * meaning “absolute, apeiron, the undifferentiated whole” that could be seen as the primordial top class of which all phenomena are subdivisions. This would have implied that all single-letter classes would have a skos:broader relationship to class *. Draft visualizations of this architecture, however, looked confusing for expected common users, as they would display a very abstract, philosophical notion with much greater evidence than classes of more common usage. We thus opted for not recording such relationship in the SKOS version of ILC2.

On the other hand, this has stimulated interesting considerations on how very general philosophical notions, such as “things in themselves” or “phenomena”, may be expressed in ILC. A provisional view, that could be implemented in ILC3, is that a top class meaning “being” can include both “absolute” that is noumena or things in themselves in philosophical terminology, and “phenomena” meant as classes of differentiated named entities, in turn including all common main classes; of these, some are “real”, that is actually existent, and can be specified by the deictic \textit{Y} already available in ILC.
5.0 Publication details

The conversion created a total of 82,534 triples describing 8,990 concepts (including 52 top concepts) and 943 properties (modelled as hierarchical specializations of skos:related), a total of 9,933 items.

URLs for individual classes and facets refer to the online schedules previously available through a PHP interface. To allow online navigation, however, these have dynamic URLs of the form http://www.iskoi.org/ilc/2/no.php?no=jUxf (for class jUxf “Bay of Fundy” taken as an example). In SKOS data, the dynamic form has been changed to a static one: http://www.iskoi.org/ilc/2/class/jUxf. This has required to set a mod_rewrite redirect instruction on the iskoi.org Apache server, so that referenced URIs are automatically converted to the dynamic form and the appropriate information is displayed.

SKOS data for ILC2 are available from http://www.iskoi.org/ilc/skos.php in Turtle, NTriples or RDF syntax.

They are also available at the BARTOC (Basel Register of Thesauri, Ontologies and Classifications) repository at http://bartoc-skosmos.unibas.ch/ILC/en/ as part of the long-term experimentation with application of ILC to BARTOC indexing (Ledl and Gnoli 2017). The SKOS version is published using Skosmos, an open source tool developed at the National Library of Finland (http://skosmos.org/). This produces various flavours of RDF output, including the commonly used NTriples format.

6.0 Visualizations

A text based visualization is available from BARTOC. Graphical displays can be created by importing the generated ILC2 NTriples RDF data file into the AllegroGraph Gruff tool. Some illustrative examples of ILC2 concepts and properties follow, where we can see the SKOS NTriples output and corresponding graph-based visualisations using Gruff and the equivalent view from BARTOC.

Note that we can observe in the SKOS output the main URI for the ILC2 SKOS scheme at iskoi.org, the concept being visualised with its preferred label (“Bay of Fundy”), its notation (jUxf), broader concepts (“Atlantic Ocean”) and their notation.

```rdfs
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix ilc2: <http://www.iskoi.org/ilc/2/class/> .

<http://www.iskoi.org/ilc/2/scheme>
  rdfs:label "Integrative Levels Classification (ILC)"@en ;
  skos:prefLabel "Integrative Levels Classification (ILC)"@en ;
a skos:ConceptScheme .

ilc2:jUxf
  rdfs:seeAlso <http://www.iskoi.org/ilc/2/details.php?no=jUxf> ;
  skos:broadert ilc2:jUx ;
  skos:notation "jUxf" ;
  rdfs:label "Bay of Fundy"@en ;
  skos:prefLabel "Bay of Fundy"@en ;
  skos:inScheme <http://www.iskoi.org/ilc/2/scheme> ;
a skos:Concept .

ilc2:jUx
  skos:notation "jUx" ;
```
Figure 1: Skosmos output and Gruff visualisations and corresponding BARTOC view for Bay of Fundy

A more elaborate example with the concept of polypteriformes shows more of a hierarchical tree. We also see an example of a descriptive Note in the Gruff visualisation.

@prefix ilc2: <http://www.iskoi.org/ilc/2/class/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

ilc2:mqvh
skos:notation "mqvh" ;
rdfs:label "ray-finned fish"@en ;
skos:prefLabel "ray-finned fish"@en ;
a skos:Concept ;
skos:narrower ilc2:jUxf .
Figure 2: Skosmos output and Gruff visualizations and corresponding BARTOC view for polypteriformes
A yet more complex example shows the variety of relationships within ILC2 and connections between concepts. For example, there are associative relationships (skos:related) between “stars” and “star clusters” and between “stars” and “plasma”.

```xml
@prefix ilc2: <http://www.iskoi.org/ilc/2/class/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

ilc2:h
  skos:notation "h" ;
  rdfs:label "celestial bodies"@en ;
  skos:prefLabel "celestial bodies"@en ;
  a skos:Concept ;
  skos:narrower ilc2:hl .

ilc2:hu
  skos:notation "hu" ;
  rdfs:label "star clusters"@en ;
  skos:prefLabel "star clusters"@en ;
  a skos:Concept ;
  skos:related ilc2:hl .

ilc2:hlb
  skos:notation "hlb" ;
  rdfs:label "subdwarf stars"@en ;
  skos:prefLabel "subdwarf stars"@en ;
  a skos:Concept ;
  skos:broader ilc2:hl .

ilc2:hlg
  skos:notation "hlg" ;
  rdfs:label "giant stars"@en ;
  skos:prefLabel "giant stars"@en ;
  a skos:Concept ;
  skos:broader ilc2:hl .

ilc2:hlj
  skos:notation "hlj" ;
  rdfs:label "supergiant stars"@en ;
  skos:prefLabel "supergiant stars"@en ;
  a skos:Concept ;
  skos:broader ilc2:hl .

ilc2:hl
  skos:scopeNote "celestial bodies where nuclear fusion occurs"@en ;
  skos:related ilc2:hu, ilc2:gf ;
  skos:inScheme <http://www.iskoi.org/ilc/2/scheme> ;
  rdfs:label "stars"@en ;
  rdfs:seeAlso <http://www.iskoi.org/ilc/2/details.php?no=hl> ;
  skos:prefLabel "stars"@en ;
  skos:notation "hl" ;
  a skos:Concept ;
  skos:broader ilc2:hl .

ilc2:gf
  skos:notation "gf" ;
```
rdfs:label "plasma"@en ;
skos:prefLabel "plasma"@en ;
a skos:Concept ;
skos:related ilc2:hl .

ilc2:hlh
  skos:notation "hlh" ;
rdfs:label "bright giant stars"@en ;
skos:prefLabel "bright giant stars"@en ;
a skos:Concept ;
skos:broader ilc2:hl .

<http://www.iskoi.org/ilc/2/scheme>
  rdfs:label "Integrative Levels Classification (ILC)"@en ;
skos:prefLabel "Integrative Levels Classification (ILC)"@en ;
a skos:ConceptScheme .

ilc2:hla
  skos:notation "hla" ;
rdfs:label "attributes of #hla"@en ;
skos:prefLabel "attributes of #hla"@en ;
a skos:Concept ;
skos:broader ilc2:hl .

ilc2:hlf
  skos:notation "hlf" ;
rdfs:label "subgiant stars"@en ;
skos:prefLabel "subgiant stars"@en ;
a skos:Concept ;
skos:broader ilc2:hl .

ilc2:hlk
  skos:notation "hlk" ;
rdfs:label "hypergiant stars"@en ;
skos:prefLabel "hypergiant stars"@en ;
a skos:Concept ;
skos:broader ilc2:hl .

ilc2:hlU
  skos:notation "hlU" ;
rdfs:label "the Sun"@en ;
skos:prefLabel "the Sun"@en ;
a skos:Concept ;
skos:broader ilc2:hl .

ilc2:hld
  skos:notation "hld" ;
rdfs:label "dwarf stars"@en ;
skos:prefLabel "dwarf stars"@en ;
a skos:Concept ;
skos:broader ilc2:hl .
Figure 3: Skosmos output and Gruff visualizations and corresponding BARTOC view for stars

7.0 Conclusion

Previous experience at the University of South Wales with the STELETO transformation tool has allowed to treat the complex syntactic structures of a freely faceted classification, such as ILC, and produce an appropriate representation of them as SKOS. On the other hand, full management of concept combinations according to ILC syntax is limited by the expressiveness of the SKOS format itself, as already discussed by Gnoli et al. (2011).

Representation of a freely faceted classification as SKOS is especially useful for the purposes of data exchange in a standard international format, making it available on the Web as linked data in view of new applications. Tools for visualization of semantic structures are another benefit of conversion to SKOS. Special applications, such as PHP scripts for navigation of ILC schedules as available on the iskoi.org website, can further exploit its expressive power.
Acknowledgments
We are grateful to Andreas Ledl for publication in BARTOC Skosmos, and to Riccardo Ridi for discussion on philosophical aspects of noumena and phenomena.

References
Abstract
The variety of indexing systems needs interoperability to satisfy global information. In mapping projects and related literature, the focus is predominantly on equivalence relationships between terms. Starting from examples of terms mapped between the Thesaurus of Nuovo Soggettario and LCSH and Rameau, the semantic relationships of terms are explored to verify correspondences and divergences in related terms. The different structure of the examined vocabularies leads to semantic networks that are not parallel. Specific and general remarks follow, in light of ISO 25964:2011-2013 and recent revisions.

1.0 Aim and background
The great variety of indexing systems, together with the use of different languages, meets the specific needs of their patrons. However, today access to information on a global scale requires interoperability, a theme widely studied, and mapping is the way to reach data coming from different sources. Many mapping tests have been carried out and a rich literature is available on the matter.1 However, the focus is predominantly, if not exclusively, on equivalence relationships between concepts/terms. This attitude creates precise maps among single nuclei. However, it ignores the semantic relationships that are traditionally and diffusely represented in controlled vocabularies, hierarchical and associative relationships. Moreover, it loses sight of the overall correspondence of the envisaged systems. What happens if we explore their semantic networks? Do we find parallel or diverging nets? Do the correspondences go on step by step or stop at the starting points?

This paper addresses this issue starting from mappings of the Thesaurus of Nuovo soggettario2 (ThNS, source vocabulary) to Library of Congress Subject Headings3 (LCSH) and to Repertoire d’autorité-matière encyclopédique et alphabétique uniifié4 (Rameau), as target vocabularies; all of them are characterized for general scope and produced by national bibliographic agencies. Some typical examples of mapping are reported, showing the equivalences and the most frequent divergences between the semantic networks and giving the starting point for justifying the divergences on the basis of the different features of the indexing languages or of their application criteria. The paper discusses some overall remarks on mappings and interoperability between different indexing languages and suggests possible alternative solutions drawn from ISO 25964:2011-2013, which is a sound reference to recognize and represent equivalences and their degree, to state meaningful and useful links where there is no equivalence, and to manage non-parallel systems.

2 https://thes.bncf.firenze.sbn.it/ricerca.php
3 http://id.loc.gov/authorities/subjects.html
**Nuovo soggettario** is an analytic-synthetic indexing language produced by the National Central Library in Florence and adopted by Italian libraries (Biblioteca nazionale centrale di Firenze 2006). Here, we are interested in the component *Thesaurus*, which is devoted to the control of vocabulary and semantic relationships between concepts. It is clearly distinct from the rules for concept analysis and strings construction, and from the authority file of assigned strings. In this separation of syntax from semantics, ThNS conforms to stipulations of ISO 25964-1:2011, and can function autonomously in post-coordinated indexing. The terms are not distinct in headings and subdivisions and there are no complex subjects. The terms included are only topic concepts, and terms that are valuable as *genre/form* are also included without distinction. Geographic terms, proper names and titles are not included. The structure of ThNS is not by discipline, but is founded on four macrocategories (Agents, Actions, Things, Time) divided into semantic categories, in number of thirteen (Organisms, Organizations, Persons and groups; Activities, Disciplines, Processes; Matter, Objects, Space, Tools, Forms; Structures; Time). Each of these categories is organized in facets and sub-facets, also adopting node labels. Within these categories, hierarchical relationships between concepts (BT and NT) and equivalence relationships between terms (USE, UF) are arranged without deviations. The connections between concepts belonging to different categories are recorded as associative relationships (RT). The use of polyhierarchy is limited to few controlled conditions.

These features make ThNS different from other widespread indexing systems like LCSH and Rameau, but do not prevent from mapping correspondences useful for searching through different systems. ThNS mapping has reached almost 14000 equivalences with LCSH and 12000 with Rameau and it is still in progress. The equivalence mappings from ThNS are all recorded in skos as ‘closeMatch’, even when the equivalence could be an ‘exactMatch’. Compound equivalence, as foreseen in ISO 25964-2:2013, 8.3, is not adopted: neither intersecting (EQ+) nor cumulative (EQ|). Mappings on broader or narrower levels (BM or NM) are not adopted, so each term admits only one mapping with one term of the same semantic category in the target vocabulary, and there are no double mappings from or to one term.

The equivalences recorded in LCSH as ‘closely matching concepts’ with ThNS are the reciprocal of the mappings made in Florence, unfortunately incomplete, as they have not been added lately. In Rameau, Italian terms do not yet appear, but in data.bnf.fr they are now included among the resources ‘*sur le Web: notice correspondante dans Le Nuovo Soggettario*’ and also among the variants (*autres formes du thème*) with the qualification ‘*italien*’. Regardless of the following technical remarks, future reciprocal agreements are essential to the effectiveness of these works: for users’ searching functions and to manage the updating through the vocabularies.

Mapping activity has shown cases of no correspondence (a concept represented in a vocabulary does not appear in the other) and of inexact correspondence (for instance, due to a different level of specificity). These mismatches can be due to language differences, different context and literary warrant, or different structure and application criteria of the considered indexing languages (for instance, the inclusion of complex subjects, requiring a double mapping from two terms of ThNS, or a different granularity). Some representative cases of difficult correspondence between single concepts have been previously reported (Buizza 2019).
2.0 Mapping concepts and exploring relationships

This study focuses on six concepts represented by equivalent terms in the three vocabularies. Their hierarchical chains and related terms have been explored, in order to verify correspondences and differences in their semantic networks. The sample of examples tries to cover the most typical and interesting cases, without any presumption to be complete. No quantitative survey was carried out, to know the incidence of unsatisfactory results.

The first example is a product globally widespread, the smartphone. ThNS records it as Smartphone (not translated), and its equivalents are Smartphones in LCSH and Smartphones in Rameau. Looking at the semantic relationships in the three vocabularies we can see immediately some differences, beginning from hierarchical ones (see Table 1). In ThNS, for the monohierarchical choice, we find only one BT Telefoni cellulari, even if the object is both a telephone and a computer. The term representing the second hierarchy, Palmari, is linked with an associative relationship, RT. In LCSH polyhierarchies are applied and we find two BTs: Cell phones and Pocket computers. In ThNS we have no NT due to the general choice of not recording proper names in the thesaurus, whether brand names or names of products. In LCSH some models from different brands are recorded as NT (for instance Samsung Galaxy S (Smartphone)). Therefore, only one equivalence can be set at the upper level (Telefoni cellulari EQ Cell phones) and no equivalence can be set at the lower level.

Table 1. The case for Smartphone (simplified); double hierarchy is split. The superscripts mark recorded closeMatches.

<table>
<thead>
<tr>
<th>ThNS</th>
<th>LCSH</th>
<th>Rameau</th>
</tr>
</thead>
<tbody>
<tr>
<td>... BT Telefoni®</td>
<td>... BT Telephone®, Radiotelefonia®</td>
<td>... BT Téléphone®, Radiotéléphonie®, Radio-</td>
</tr>
<tr>
<td>RT Radiotelefonia®</td>
<td>UF Telephone service, Telephones</td>
<td>communications mobiles, Systèmes de</td>
</tr>
<tr>
<td>· BT Radiotelefono®</td>
<td>· BT Cell phones²,³</td>
<td>communication sans fil</td>
</tr>
<tr>
<td>· BT Telefoni cellulari²</td>
<td>· NT Samsung Galaxy S (Smartphone)⁶</td>
<td>· BT Téléphonie mobile®</td>
</tr>
<tr>
<td>Smartphone¹</td>
<td>· BT Portable computers®,⁸</td>
<td>· EP Téléphones cellulaires</td>
</tr>
<tr>
<td></td>
<td>· BT Poket computer®²</td>
<td>· NT Samsung Galaxy (Smartphones)³</td>
</tr>
<tr>
<td></td>
<td></td>
<td>· NT Samsung Galaxy S (Smartphone)³</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>· BT Ordinateurs portatifs®,⁸</td>
</tr>
<tr>
<td></td>
<td></td>
<td>· BT Ordinateurs de poche²</td>
</tr>
<tr>
<td></td>
<td></td>
<td>· Smartphones¹</td>
</tr>
</tbody>
</table>

The comparison with Rameau gives similar results. In Rameau, two BTs are recorded, Téléphonie mobile and Ordinateurs de poche, and some NTs for specific brands, with relationships parallel to those of LCSH. However, there are two differences. The BT Téléphonie mobile has shifted the semantic category from objects to an activity – or rather, concepts of activity and of objects have been merged under one term, as shown by the French equivalents in Rameau (e.g. EP Téléphones cellulaires). Based on the category distinction, ThNS cannot set an equivalence with this term as it would cover only the component ‘object’. The parallelism is lost also with LCSH and Téléphonie mobile has correspondence with both Cell phones and Cell phone systems, with two distinct relationships. The second difference is how the two heading lists link the proper names of the smartphones to the noun term. LCSH links them directly as NT. Rameau
links them in two steps, through an intermediate term for the brand: Smartphones NT Samsung Galaxy (Smartphones) NT Samsung Galaxy S (Smartphone). The last term has as equivalent the homograph recorded in LCSH, while the intermediate term has no equivalent in LCSH.

The ways hierarchical chains go to upper levels are now explored. In ThNS Telefoni cellulari has BT Radiotelefoni, which has BT Telefoni. In LCSH Cell phones has BT Telephone (in the singular, with variants: Telephone service and Telephones, with double category value) and also the complex subject Radio—Transmitters-receivers. ThNS sets equivalence between Telefoni and Telephone, even though the Italian word does not include the telephone service. No equivalence for Radiotelefoni, but the RT Radiotelefonia has equivalent Radiotelephone, that includes service and devices (its NT Walkie-talkies has a homograph in ThNS recorded as NT of Radiotelefonì).

In Rameau, Téléphonie mobile has four BTs: Téléphone, Radiotéléphonie, Radiocommunications mobiles and Systèmes de communication sans fil, but the last term has the second and third as NT, besides Téléphonie mobile, and the second term is also NT of the first; an entry to be updated, probably – there are some in every vocabulary. The equivalences in ThNS go from Telefoni and Radiotelefonia to Téléphone and Radiotéléphonie – equally from the English terms to the French. The correspondences between the three vocabularies are restored, but the paths from smartphones to this point are different and the exact correspondence between smartphones is not kept between the reference to one category in ThNS and the meaning merging object and functions in LCSH and Rameau.

According to the choice of monohierarchy, ThNS establishes an associative relationship to the second hierarchy: Smartphone RT Palmari. The equivalents of Palmari are Pocket computers and Ordinateurs de poche. Their respective BTs, Computer portatili, Portable computers, and Ordinateurs portatifs, are equivalent. At upper levels we find the equivalence of the three languages. But in LCSH and Rameau coming back to smartphones is inside one hierarchical chain by NT relationships, while in ThNS it is necessary to pass through the associative relationship Palmari RT Smartphone.

The complexity and the divergencies of these semantic networks are due, at least in part, to the double function of the considered object. Another example is the simple concept of ‘date’, the fruit, seen from different points of view. In ThNS the term Datteri has BT Frutta, whose BT is Alimenti, and RT Frutti, a term that has BT the node label [Organi e parti di piante], that has BT Piante. According to the meaning of the Italian words, Frutti has a scope note assigning it the botanic meaning, while for the works about fruits as food Frutta is used. In LCSH, the equivalent Dates (Fruit) has two BTs: Date palm products (with hierarchy: BT Palm products BT Plant products) for the economic side, and BT Fruit, for food, agricultural and botanic meanings, as is shown by its five BTs: Food, Food crops, Horticultural crops, Horticultural products, and Plants. In Rameau, the equivalent Dattes has BT Fruit, whose scope note states both botanic and food meanings. These are shown together with the economic meaning, in its three BTs: Diaspores (botanique), Plantes comestibles, and Produits horticoles. In LCSH and in Rameau, we find also Cooking (Dates) and Cuisine (dattes), with the activity of cooking qualified by the specific food, a method collecting a long sequence of terms for each type of food under the same first word. These complex terms cannot have correspondence in Nuovo soggettario, where strings are provided syntactically out of the thesaurus.
In this case the given concept is represented by the term for the food and a form term meaning the aim of the document: *Datteri-Ricette*. This example confirms that: adopting polyhierarchy or not is an important factor; scope notes may be crucial for stating an equivalence; the presence of complex subjects makes a difference between the vocabularies that is difficult to manage.

Even if there are not different meanings, nor different contexts for a concept, the equivalence of the terms for a given concept does not grant the persistence of the equivalence in the other steps of the semantic networks.

Looking at concepts representing a set of individuals, for instance deities, and considering a particular historical expression, say Greek deities, we find the equivalent terms *Divinità greche*, *Gods*, *Greek*, and *Dieux grecs*. In ThNS the subordinate terms are two subsets, *Muse* and *Ninfe*, while in LCSH we find some subsets of different kind (e.g. *Gods*, *Minoan*, referring to specific areas) and a number of proper names (e.g. *Chaos* (*Greek deity*)) but not the well-known gods of Olympus (recorded in Library of Congress Name Authority File). In Rameau we find some subsets (e.g. *Dieux minoens*), the female collective term (*Deesses grecques*) and an exhaustive list of proper names of deities (e.g. *Aphrodite* (*divinité grecque*)). The choice of *Nuovo soggettario* to manage the proper names of individuals through the guidelines of the *Manuale*, causes this lonely position of ThNS. However, the other vocabularies do not agree with each other either: the distribution of subordinate sets is not equal and LCSH places some gods among names of persons.

There are misalignments in superordinate relationships too. The Italian term has BT *Divinità*, the English one lacks BT, the French one has two study areas as BT *Mythologie grecque* and *Religion grecque*. The equivalent terms for *Divinità* are *Gods* and *Dieux*. *Divinità* has NTs for deities of various peoples and *Divinità femminili*, *Divinità marine*, *Divinità salutari*, etc. *Gods* has NTs for deities of nature (e.g. *Water gods*) and of specific religions (e.g. *Hindu gods*, with NT for some proper names), except for classical religions, that are orphan, and has BT *Mythology: Classical*. *Dieux* has NTs for the deities of phenomena (e.g. *Dieux des vents*), that can have NTs for proper names, like *Vayu* (*divinité hindou*), while the terms qualified by people or religion have BT for the appropriate mythology and/or religion (e.g. *Dieux hindous* BT *Hindouisme, Mythologie hindoue*).

Other differences in semantic relationships can be found in the category of persons. *Donatori di sangue* has equivalents *Blood donors* and *Donneurs de sang* (see Table 2). The term in ThNS has BT the node label *[Persone secondo il comportamento]*, that has BT *Persone*. In LCSH the immediate BT is *Persons*. This is not a small difference: *Persone* has three node labels (according to activity and to conditions, in addition to behaviour) and each of them has NTs for other node labels in a widely faceted articulation. Under *Persons* a long list of terms represents specific types of persons (e.g. *Saints*, *Slaves*, *Travelers*), arranged alphabetically. In a different way, in Rameau *Donneurs de sang* has BT *Donneurs d’organes*, that has no BT, the same as the equivalent terms of the above examples: *Saints, Esclaves, Voyageurs*. As a result, the terms for categories of persons are fragmented and neither systematized in a pyramid as in ThNS, nor collected under a comprehensive top term as in LCSH. There are some groupings, for instance under *Catégories socio-professionnelles* (with hierarchies like: NT *Commerçants* NT...
Libraires), and also terms with a BT. But some BTs might have moved to another semantic category, for instance, Personnes remariées has BT Remariage.

Table 2. The case for Blood donors and Persons. The superscripts mark recorded closeMatches.

<table>
<thead>
<tr>
<th>ThNS</th>
<th>LCSH</th>
<th>Rameau</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persone$^2$</td>
<td>Persons$^2$</td>
<td>--</td>
</tr>
<tr>
<td>· NT [Secondo l’attività]</td>
<td>· NT Slaves$^3$</td>
<td>--</td>
</tr>
<tr>
<td>· NT [Secondo la condizione]</td>
<td>· NT Organ donors$^5$</td>
<td>--</td>
</tr>
<tr>
<td>· NT Schiavi$^3$</td>
<td>· NT Blood donors$^1$</td>
<td>· NT Donneurs de sang$^1$</td>
</tr>
<tr>
<td>· NT [Secondo il comportamento]</td>
<td>· NT Travelers$^4$</td>
<td>· NT Vaiaggiatori$^4$</td>
</tr>
<tr>
<td>· NT Donatori di organi$^6$</td>
<td>· NT Saints$^5$</td>
<td>· NT Sant$^5$</td>
</tr>
<tr>
<td>· NT Viaggiatori$^4$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>· NT [Secondo la fede e le convinzioni religiose]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>· NT Santi$^5$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A term for a discipline, Dermatologia, has equivalents Dermatology and Dermatologie (see Table 3). In ThNS the broader term Medicina is reached through the interposed node label [Medicina applicata a specifici organi, apparati, sistemi, funzioni]. This node label is paired on the same array by two other labels: [Medicina applicata a categorie di persone] and [Medicina applicata a specifiche attività], which are all recorded as NTs under Medicina specialistica. The result is a faceted distribution of the branches of medicine. In LCSH and Rameau there is a direct relationship to the general discipline, BT Medicine and BT Médicine respectively. In the subordinate hierarchy, in ThNS there are no NTs and the term Dermatologia veterinaria is recorded as associated (RT). In a different interpretation, its equivalents Veterinary dermatology and Dermatologie vétérinaire are recorded as NT. Among other NTs for specialties there are also terms that are neither members nor parts of the superordinate concept, such as, in LCSH, agents: Dermatologists, and a technique: Radioisotopes in dermatology, or, in Rameau, a védette construite for an activity: Peau-Maladies-Soins infirmiers.

Table 3. The case for Dermatology. The superscripts mark recorded closeMatches.

<table>
<thead>
<tr>
<th>ThNS</th>
<th>LCSH</th>
<th>Rameau</th>
</tr>
</thead>
<tbody>
<tr>
<td>· BT Medicina$^2$</td>
<td>· BT Medicine$^2$</td>
<td>· BT Médecine$^2$</td>
</tr>
<tr>
<td>· BT Medicina specialistica</td>
<td></td>
<td></td>
</tr>
<tr>
<td>· BT [Medicina applicata a organi...]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dermatologia$^1$</td>
<td>Dermatologie$^1$</td>
<td>Dermatologie$^1$</td>
</tr>
<tr>
<td>· RT Dermatologia veterinaria$^3$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>· NT Veterinary dermatology$^3$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>· NT Pediatric dermatology$^4$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>· NT Dermatologists</td>
<td>· NT Dermatologie vétérinaire$^3$</td>
<td></td>
</tr>
<tr>
<td>· NT Radioisotopes in dermatology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>· NT Peau-Maladies-Soins infirmiers</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Another disciplinary term, Diritto internazionale, and its equivalents International law and Droit international, present the same kind of difference at superordinate levels: direct relationships to Law and Droit, while in ThNS a node label is interposed: BT [Diritto secondo la materia] BT Diritto. The subordinate Italian hierarchy has specific branches (e.g. Diritto comunitario, Diritto internazionale marittimo), and associative relationships link objects, tools and activities of the discipline. In LCSH and Rameau
there are many NTs, including any kind of concept that can be considered in the semantic area of the discipline, including complex or syntactically constructed terms (e.g. Islands-Law and legislation or Iles-Droit) and parenthetic terms (e.g. Missing persons (International law) and Personnes disparues (droit international)). The sets of the two languages do not fully overlap, some terms have no equivalent in the other language (e.g. Women (International law) or Animaux (droit international)).

In ThNS a parenthetic term is allowed only for disambiguation of homographs (normally not by the discipline). Concepts expressed with parenthetic terms in LCSH and Rameau are seen in ThNS as syntactical relationships between distinct concepts, to be treated according to the rules for string construction (if they are not pleonastic). The term Intervento (Diritto internazionale), with the same meaning of Intervention (International law) and of Intervention (droit international), is a non-preferred term to be expressed by a combination of terms: Intervento (Diritto internazionale) USE+ Diritto internazionale, Intervento militare. No equivalence is possible from the Italian terms to those of LCSH and Rameau. According to ISO 25964-2:2013, 8.3.2, an intersecting compound equivalence could be provided in the other direction: Intervention (International law) EQ Intervento militare + Diritto internazionale, and Intervention (droit international) EQ Intervento militare + Diritto internazionale. Obviously, this method is not adopted, at present.

In ThNS Discipline is a semantic category and a top term. Its hierarchy comprises only concepts of discipline. Other concepts strictly related to a discipline (not simply falling into a discipline) are linked with an associative relationship, and remain in their semantic category (agents, activities, tools, etc.). Thus, in ThNS we do not find the hierarchies typical of classifications based on disciplines, where a discipline is the systematic container of everything it concerns. The free structure of LCSH and Rameau is quite different: categories are not considered and the relationships with narrower terms may be similar to an alphabetic index collecting all the topics belonging or attributed to that discipline, including complex and parenthetical terms.

3.0 Overall remarks

There are several differences between the semantic networks of the three vocabularies, even though we can often start from equivalent terms for the same concept and we find other full equivalences along the paths. The differences occur mainly between ThNS and the two heading lists, due to the different structure of ThNS explained above. To summarize, some of the reasons are: firstly the lack of subdivisions, complex subjects, and proper names. Secondly, an architecture based on categories without disciplinary groupings, strictly categorial equivalences and hierarchies, faceting, and limited polyhierarchy, compared with more free relationships aiming to link what is inside a discourse rather than the accurate identities of concepts. Téléphonie mobile is not the same entity as Téléphones cellulaires, but when talking about the former one talks also about the latter.

There are important differences even between the strongly connected LCSH and Rameau, due to different granularity or definition, including jumping hierarchical levels, merging concepts under one term or distinguishing them under different terms, different criteria of subdivision, non-identical choices in recording complex subjects. This means
that some equivalences are missing or couples of terms referred to the same concepts need different paths to be connected.

The great extension of the vocabularies allows for the presence of most concepts in the three languages and their mapping, so the divergences of the networks may appear not very important: surfing among terms seems always possible. What are the drawbacks of these divergences, in regard to the primary value of the equivalence between terms representing the same concept?

Terms have no magic power. They can neither fully represent the content and informative potential of the works entered under them, nor fully and exactly retrieve the searched information. They live as strictly connected elements of indexing and of natural language, controlled in indexing and free in users’ access. Terms literally say what their definitions or scope notes say, but they represent wider themes and thoughts, multiple and articulated connections to other concepts and themes. Semantic relationships just try to represent these articulated meanings in systematic, understandable and viable order, guiding indexers in choosing terms and users in searching and exploring resources. Different configurations of the semantic frames are quite legitimate and sometimes required. However, for interoperability and combined use, the largest overlapping of the systems, or, at least, knowing the different paths and the missing links is desirable. Going up and down hierarchies consistently makes the choice of broadening or narrowing the field of research easier, faster and more reliable, and avoids possible confusions. The same usefulness works in order to see the sibling terms, passing through the broader term.

In indexing work, only the adopted system is involved and the problems of semantic networks are less relevant than the deficiencies in mapping. Some observations about the mentioned cases reduce their seriousness. The lack of proper names in ThNS simply means that a name is assigned following the Manuale and is recorded and searchable among the strings, not in the thesaurus. Complex subjects are ready to use, while in ThNS the strings are combined when needed, according to syntactic rules, or drawn from strings already recorded. The equivalence is restored at this level. Missing terms for a concept, probably due to lack of literary warrant, could be added and validated (e.g. Divinità minoiche in ThNS, or Samsung Galaxy smartphones in LCSH), unless an equivalence relationship with a term for a slightly different concept is preferred in the vocabulary policies. Only when two concepts are represented by the same term we can find different solutions in indexing. Obviously other differences in indexing results come from different policies and syntax (e.g. summarization vs depth indexing, or one coextensive string for one work vs more complementary strings).

The results of searching through diverging semantic networks vary greatly depending upon the way indexes are presented and searches are carried out: through the vocabulary or directly on strings, by surfing the vocabulary or the strings or the records, with or without automatic extension to target vocabularies. ‘Exploding’ a search, that is, adopting a technique able to retrieve also the resources linked to the terms subordinate to the searched term, gives different results in systems with or without polyhierarchies, and adopting hierarchies with or without extra semantic categories. The same happens ‘expanding’ a search, that is, retrieving the resources linked to both subordinate and associated terms (ISO 25964-1:2011, 10.2.1). In general terms, simple searches are not problematic. When surfing, users must follow the different features of each indexing system.
Automatic exploring gives variously inaccurate results (with topics and resources that are not pertinent), insufficient (some relevant resources are not found) or redundant (relevant resources are retrieved together with off-topic ones, good for serendipitous diversions), depending on finding equivalent, lacking or overabundant relationships.

Nowadays the connections between the equivalent terms of the different vocabularies allow users to search in the catalogues of the three libraries in few steps: any improvement of interoperability will be beneficial to users from the three areas. Structural differences and soundness of the systems seem unfavourable to future convergences. However, some suggestions are an obvious consequence of this survey. Adding other forms and levels of mapping is recommendable to clarify and enrich connections, even though it acts almost only at the level of single correspondences, without weakening divergencies between the semantic networks. Any convergences towards international standards, especially ISO 25964, would lead to positive effects, even though subject headings lists are too far from thesauri in their structure.

At present, some elements offer suitable conditions to interoperability. Linked open data already support the three systems and work on data of the entities in a way that prefers modular systems, instead of systems fixing pre-coordinated entities in their vocabularies. Systems where each element is treated per se, in order to be combined with other elements, either steadily according to semantic paradigms or occasionally according to syntactic needs, are more suitable for linked data. In this perspective, ‘Réformer Rameau’ is a very encouraging venture launched in France\(^5\). In five years, among other things, it foresees to abolish the distinction between tête de vedette and subdivision, to split up the vedettes construites, and to delete special instructions for disciplines. This treatment of terms is similar to thesaural treatment. It is neutral as regards the destination to pre-coordination, and reduces the disciplinary bonds, that functionally collect any object of interest in a given field of research, but constrain concepts into limits more restricted than those where concepts normally move.

As for ThNS, the second edition of the Guida to Nuovo soggettario, now in the final phase, confirms the adherence to standard ISO 25964. It does not change the features of the thesaurus and its criteria for mapping other vocabularies, even after a complete rewriting. We cannot exclude a future opening to varied mappings, which would be particularly helpful if accompanied by collaboration and agreed criteria between national libraries.

What matters here, in this description of coherence, limits and problems of network mapping, is to point out that good mappings between single terms are not sufficient. This paper recommends, particularly to people involved in designing interoperability, going beyond mappings between single terms, towards clearly branched and connected maps, which would be useful to cross unknown or less well-known areas of knowledge.

References


\(^5\) https://rameau.bnf.fr/syntaxe


Inheritance and Lamination in the Representation of Bibliographic Relationships

Abstract:
This paper uses the concept of lamination implied in the term “discovery layer” to explore how domain knowledge could be applied to large federated search environments. Using the publishing history of Daniel Defoe’s *Robinson Crusoe* as a case study, we use bibliographic scholarship in the English literary studies community to establish lines of inheritance on all four levels of the FRBR paradigm: work, expression, manifestation and item. We created a small demonstration of a visualization generated from linked data extracted from the scholarly literature, to show how literary scholarship, when encoded as linked data, can create lines of inheritance and influence that enable users to fulfil the fifth user task of the new Library Reference Model: exploration.

1.0 Introduction

Knowledge organization, as a field of both theory and practice, regularly grapples with the place of specialized domain knowledge within information systems that purport to be “universal” in breadth and use. While the teams that maintain the Decimal Classification, the Universal Decimal Classification and the Library of Congress Classification rely on subject experts to help them maintain various sections of these tools, the relationship between domain knowledge on the one hand, and principles of knowledge organization that purport to be a priori assumptions on the other, is by no means clear (Hjorland 2015). In this paper, we address this conflict in a different way by focusing on bibliographic rather than thesaural relationships: in particular, we examine the patterns by which one bibliographic entity influences others over time. By considering the possibilities of using linked data to encode relationships across federated information resources, we explore new ways in which specialized domain knowledge can be integrated into library catalogues.

2.0 Federated Searching and the Concept of Lamination

The concept of federated searching—simultaneously searching multiple information sources—has a long history in librarianship, dating back at least to the National Union Catalogue of the Library of Congress. Efforts to establish global standards of information description—“Universal Bibliographic Control”—rest at least partly on the dream of searching diverse catalogues through the standardization of description methods. On the Web, metasearch engines and metadata harvesting initiatives have led to the call of Tim Berners-Lee for data that is freed from its local context and permitted to combine in productive and exciting ways (Berners-Lee 2009).

The concept of the “discovery layer” has signalled a new environment in which federated searching in libraries has gone to a new level of complexity and power. Discovery layers are search platforms that integrate online catalogue data with data from other sources, enabling libraries not only to provide access to a wide range of resources through a single interface or portal, but also to customize search results in specialized ways (Ramsay & Chamberlain 2012). Not only do these discovery layers permit
searching across a range of different catalogues; they can also create interfaces that mimic popular search engine interfaces, permit faceted filtering of search results, and allow enhanced means of user feedback and participation in community search practices.

Much of the challenge with discovery layers lies, naturally enough, with problems of data harmonization: how to enable metadata structured and encoded for one resource to combine with metadata from a different resource entirely. The promise of linked data, combined with the emergence of the Dublin Core as a lingua franca to which multiple diverse systems can convert, has made such harmonization possible, producing interfaces such as Western Libraries’ new federated search interface, OMNI, which allows users to search 14 Ontario university libraries for books, articles, videos, music and databases using one common interface.

The term “discovery layer,” however, evokes another mental image as well. While proprietary platforms such as ProQuest’s Summon offer limited opportunities for customization, the rise of open source platforms such as VuFind and Blacklight offer greater opportunities for innovative adaptation (Barber, Holden & Mayo 2016, 182). Such discovery layers make it possible for users to add their own tags to records, creating folksonomies that are effectively “laminated” over the library catalogue, creating bibliographic relationships of specific utility. The concept of lamination—the superimposition of multiple data layers upon a foundational piece of data—was invoked by Ranganathan to describe the process of adding isolates to a base number to create a compound subject (Ranganathan 1967, 354). Later, the rise of geographic information systems popularized the process of overlaying geographical areas with successive layers of data indicating different facets of geographical interest (Hawkins 1994, 94).

This paper describes a case study which explores the feasibility of incorporating specialized domain knowledge into a linked data discovery layer that could be laminated over bibliographic data—particularly union catalogues—in order to encode and visualize bibliographic relationships in fresh ways.

3.0 Bibliographic Relationships

Much of the development in bibliographic description since the heyday of the Anglo-American Cataloguing Rules has involved the development of richer and more varied bibliographic relationships. The digital environment has enabled us to move beyond those relationships defined in the Paris Principles and which had to be primarily implied through main and added entries that determined where a record appeared (or did not appear) in a library’s card catalogue. The paradigm of the Functional Requirements of Bibliographic Records (FRBR), which serves as the basis for AACR2’s replacement, Resource Description and Access (RDA), separates a resource into four distinct entities, Work, Expression, Manifestation and Item, each of which exists in a one-to-many relationship with its predecessor. This paradigm is exploited in RDA to facilitate four primary objectives: finding resources, selecting them, identifying them and obtaining them (RDA 0.4.2.1) In addition, RDA provides a set of relationship designators for each level of the paradigm that can be used to encode relationships between different resources, such as abridgments, adaptations, digests, inspirations, remakes and variations (RDA Appendix J).
With IFLA’s recently-released Library Reference Model (LRM), library catalogue standards have admitted a new objective. In addition to existing objectives to enable the user to find, identify, select and obtain, the LRM has added “Explore” as a fifth important user task: “to discover resources using the relationships between them and thus place the resource in context” (Riva, Le Boeuf & Žumer 2017, 15, emphasis added).

The question lurking behind this new functionality is, of course: who is going to do all of this extra encoding? One option would be descriptive cataloguers themselves: the existing relationship designators would enable cataloguers to do a certain amount of it, embedding the relationships directly into the bibliographic records. In stark contrast, many discovery layers suggest another option by including user tagging affordances in their features, enabling library users to recognize and flag relationships of meaning to themselves.

This paper investigates a middle approach between these two extremes: the use of professional domain knowledge encoded to create bibliographic relationships between different resources: relationships that can exist at various levels of the FRBR paradigm. In this particular case, we are examining relationships defined in the field of literary studies, drawing on published literary and bibliographic scholarship to define relationships that could be encoded using linked data that refers to existing bibliographic and authority records, thereby enable users to follow temporal progressions of influence and development in literary history. Patterns of influence and inheritance would be effectively “laminated” over the bibliographic data when needed.

4.0 Temporal Progression

The use of temporal progression as a guiding principle in this case study requires some justification: not all literary studies scholarship is rigidly defined by chronology. Nonetheless, a significant amount of literary scholarship analyzes patterns of influence, both documented and implied, patterns of development of certain genres and themes, and verification of the accuracy of certain statements. Time therefore figures prominently in many literary studies, particularly studies of literary bibliography.

Classification theorists acknowledge the omnipresence and importance of time as a facet, and Fairthorne anticipated a growing awareness, within and beyond information science, that important relationships of inheritance can only be understood through a close attention to the temporal aspects of classification, whether natural or bibliographical (Fairthorne 1985, 363). More recently, scholars have acknowledged a shift in scientific thinking away from Linnaean taxonomies to a cladistic paradigm that recognizes lines of descent across time, thereby redefining category membership in ways that Linnaean classification does not allow (Hjorland 2015).

While this case study does not adopt a formally cladistic approach, we proceed from the assumption that charting scholarship in literary studies and literary bibliography along a temporal dimension provides a fruitful means of overlaying domain expertise upon a large information store for purposes of bibliographic exploration.

5.0 The Case of Robinson Crusoe

Daniel Defoe (1661?-1731) is an author who cries out for scholarly assistance in the act of bibliographic control. An extraordinarily prolific author, Defoe wrote voluminously on a dizzying range of subjects, at a time when authorial attribution was
highly unreliable. As a result, the number of works attributed to Defoe has grown over the years, only to be cut back ruthlessly in the 1990s when Furbank and Owens aggressively refuted many of the attributions (1998). As a result, conventional bibliographic control can be perilous in the case of Defoe: The Life of Mrs. Christian Davis, a sensational history of a cross-dressing Irish matriarch who distinguished herself as a foot soldier before succumbing to dropsy and scurvy, was mistakenly attributed to Defoe for some time, and the bibliographic metadata of at least one published version reflects this mistake.

Robinson Crusoe lends itself particularly to scholarly assistance. It was “a milestone in literary history” that was immediately embraced (and sometimes ridiculed) well beyond England (Backscheider 1989, 412). What is more, the novel exhibits complexities, both in its publishing and its influence, that touch on all four levels of the FRBR paradigm. Defoe himself wrote two sequels: The Further Adventures of Robinson Crusoe later in 1719, and Serious Reflections During the Life and Surprising Adventures of Robinson Crusoe in 1720. There were many translations of the novel into other languages, as well as multiple abridgments, some unauthorized. The novel’s premise of a man shipwrecked on an island and growing in wisdom and knowledge inspired a long tradition of adaptations known as “Robinsonades,” some for adults and many for children.

In exploring these complexities, we envisioned a federated search environment that would include the catalogues of numerous academic libraries, together with their extensive authority records, and such databases as the English Short Title Catalogue. Using authoritative scholarly and bibliographic guides to the publishing history of Robinson Crusoe, we began tracing inheritances at all four levels of the FRBR paradigm: Work, Expression, Manifestation and Item.

5.1 Work Relationships

Robinson Crusoe was first published in 1719. The Library of Congress Name Authority File contains a name-title authority record for the work:

Defoe, Daniel, 1661?-1731. Robinson Crusoe
LC Authority: https://lccn.loc.gov/n81045585
Publication date: 1719

Using the Name Authority File together with scholarly data, we can isolate at least two lines of influence at the Work level: sequels and Robinsonades.

5.1.1 Sequels.

Hutchins (1925), Lovett (1991) and Furbank and Owens (1998) all agree that Defoe published two sequels to Robinson Crusoe, one in 1719 and one in 1720. The LC Authority File contains name-title access points for both:

Defoe, Daniel, 1661?-1731. Farther adventures of Robinson Crusoe
LC Authority: https://lccn.loc.gov/n94091299
Publication date: 1719

Defoe, Daniel, 1661?-1731. Serious reflections during the life and surprising adventures of Robinson Crusoe.
LC Authority: https://lccn.loc.gov/n94091314
5.1.2 Robinsonades.

Searching the MLA International Bibliography using the term “Robinsonades” produces a number of scholarly articles that identify works drawing on Robinson Crusoe. Among those with LC Name-Title Authority Records are:

- Dalayrac, N. (Nicolas), 1753-1809. Azémia
  LC Authority: https://lccn.loc.gov/no93006886
  First Performed: 1786
  A French comic opera

- Ducray-Duminil, M. (François Guillaume), 1761-1819. Lolotte et Fanfan.
  English
  LC Authority for English Translation: https://lccn.loc.gov/n00075009
  First published: 1788
  Novel, known in England as Ambrose and Eleanor

- Campe, Joachim Heinrich, 1746-1818. Robinson der Jüngere
  LC Authority: https://lccn.loc.gov/n95060785
  First published: 1788
  Instruction book for children

  Polish
  LC Authority: https://lccn.loc.gov/no2015153207
  First published: 1915
  Novel

- Golding, William, 1911-1993. Lord of the flies
  LC Authority: https://lccn.loc.gov/no2006021893
  First published: 1954
  Novel

5.2 Expression Relationships
5.2.1 Translations

Robinson Crusoe was translated into many languages. A recent exhibition at Indiana University’s Lilly Library (http://www.indiana.edu/~liblilly/defoe/translations.html) provides the following examples of early translations:

First French Translation, 1720:

Italian Translation, 1731:
Translation based on the French version above.

Second Dutch Translation, 1735:
Het Leven en de wonderbare Lotgevallen van Robinson Crusoe, behelzende onder andere ongehoorde uitkomsten een verhaal van zijn acht-en-twintigjaarig verblijf op een onbewoond eiland, gelegen op de kust van America, bij de mond van de rivier Oronoouque. Alles door hemzelf beschreven. Door G. Schreuders. Amsterdam, 1735-1736.

Arabic Translation, 1835:
Qi•s•sah R¯ubin•sun Kr¯uz¯i.Malta, 1835.

Persian Translation, 1878:

5.2 Abridgments
At least three abridgments of Robinson Crusoe appear in the English Short Title Catalogue:

The Midwinter Abridgment, 1722:
The life and most surprizing adventures of Robinson Crusoe, of York, mariner. Who lived eight and twenty years in an uninhabited island on the [co]ast of America, lying near the mouth [of] the great river of Oroonoque: ... The whole three volumes faithfully abridged, ..
English Short Title Catalogue (ESTC) Number: 006343293

An Abridgment of the Midwinter Abridgment, 1734:
The wonderful life, and most surprizing adventures of Robinson Crusoe, of York, mariner ... Faithfully epitomized from the three volumes, and adorned with cutts suited to the most remarkable stories .. printed for A. Bettesworth and C. Hitch, at the Red-Lyon; and J. Osborn, at the Golden-Ball in Pater-Noster-Row; R. Ware in Amen-Corner, and J. Hodges at the Looking-Glass on London-bridge, 1734.
ESTC Number: 066477403
Abridgment, 1790:
A concise abstract of the wonderful life, and surprising adventures of that renowned hero, Robinson Crusoe, who lived twenty-eight years on an unhabited island, and was afterwards released by pirates. Adorned with cuts. London: printed for, and sold by all the stationary and toy shops in town and country, [1790?]
ESTC Number: 006061120

5.3 Manifestation Relationships
Manifestation relationships have a significant potential to assist scholars in eighteenth-century literary studies by encoding important textual decisions made by modern editors. Modern literary studies frequently establish “standard editions” of canonical authors: editions which scholars prefer to consult and cite, and which often prove the basis for subsequent editions.

In the case of Robinson Crusoe, the de facto “standard edition” is the 2009 edition of the novel which appears in the Chatto and Windus complete edition of Defoe Novels. This edition was created by consulting Defoe’s first edition together with Defoe’s manuscript (Owens 326). But other lines occur as well. The first edition of 1719 was reprinted in 1827 as the Shakespeare Head Press edition, which served as the base text for the Norton Critical Edition, published in 1975 (Shinagel 1994). The Broadview Press edition of Robinson Crusoe studied a handful of editions published in Defoe’s lifetime (Davis 2014, 37).

In many instances in literary history, significant differences exist between early editions, such as the difference between the first and fourth editions of Samuel Richardson’s Clarissa, or the 1799, 1805 and 1850 editions of Wordsworth’s The Prelude. In such cases scholars and instructors, when selecting a text for use or for teaching, would very much want to trace modern editions that are based on a particular earlier edition.

5.4 Item Relationships
In some cases, textual decisions are made, not just on the basis of a particular edition, but on a particular copy of an edition: one which contains perhaps an author’s handwritten marginal notes, or which contains a half-sheet imposition correcting a compositor’s error. In the case of Robinson Crusoe, an editor of a modern edition, skeptical of the vagaries of text reproduction in the eighteenth and nineteenth centuries, might prefer to overlook the Shakespeare Head reprint of 1827 and go directly to the copy of the first edition held in the British Museum.

6.0 Visualization
As a preliminary experiment in how these relationships might be visualized, we created a sample visualization. While this visualization was created manually, the later steps of actually rendering the image can potentially be automated. The visualization and its creation thus serve as a proof of concept for a future discovery layer.
To create this visualization we created a dedicated ‘catalogue wiki’ using the MediaWiki software along with the Semantic MediaWiki extension. Next, we gathered the access points for a sample of works referred to in scholarly articles as Robinsonades:
we used the name-title access points in the Library of Congress Name Authority File. This information was then added to the catalogue wiki, creating simplified authority records in a wiki environment.

We then added to the catalogue wiki a ‘SourceOf’ relationship from Robinson Crusoe to the Robinsonades group. The Semantic MediaWiki extension allows for the encoding of these sorts of semantic web linkages within the MediaWiki environment.

The following steps to create the visualization were done manually. We exported the authority records in a standardized RDF format to visualization software. In this case Gephi was used as it has a web RDF import extension that is easily enabled. Once the RDF information was imported into Gephi extraneous information was removed and labels were simplified for human readability. One of the data points preserved on all records was the date of the work’s creation, allowing the records to be arranged chronologically.

The final result is an image that is derived from the content of the authority records (see Fig. 1). With the addition of the authoritative relationships, the final image contains some elements reminiscent of a family tree, and other elements reminiscent of a cladogram.

![Image of Robinsonade Descendents of Robinson Crusoe](image)

**Figure 1: Robinsonade Descendents of Robinson Crusoe.**

### 7.0. Conclusion

Most of these bibliographic relationships can be encoded by cataloguers in existing MARC records using the relationship designators provided by RDA. However, the task would be overwhelming. A laminated set of linked data relationships established by knowledgeable domain experts, however, could conceivably be employed to draw certain paths through a large federated collection of bibliographic records and authority files. Such paths would facilitate the new objective of “Exploration” advocated by the Library Reference Model, and also support a variety of scholarly approaches to a specific knowledge domain.
References
Ethical Perspective on Classifications of Religions
The Protestant Rise in Brazil

Abstract
This paper aims to make a comparative review on the classification of religion of the Brazilian Census in 2010, with a specific attention on the protestant groups and denominations. Classifying religions is an arduous task and there is no consensus on the best way to classify them. In fact, there is not even consensus on what differs religion, denomination, sect or cult. Clearly there are ethical issues when the classification seeks to hierarchize or make one given religion more “developed” than another or simply the action of concealment of certain religions within the category "Others". Our analysis starts from the usual bibliographic classification schemes (CS) such as Universal Decimal Classification (UDC) and the Dewey Decimal Classification (DDC), but it also seeks to investigate the used schemes in China, India and other nations with a huge population but not aligned with the Western cultures in the global north. We point the need of ethical perspectives of classification and the necessity to be careful in order to respect cultural issues and not to allow any kind of prejudice, especially in the religious aspect and in the self-determination of religious minorities who do not accept or understand certain categorizations of their beliefs. The case of the 2010 Census is then detailed, bringing with it the context where there has been continued growth in Protestantism since 1970 in Brazil and their steadily rising influence on politics. It also shows the difficulty in determining which groups make up the Protestants according to the methodological choices of the institution that conducts the Census (IBGE).

1.0 Introduction
This paper is part of a doctoral research in Information Science, in progress, over political campaign funding and protestant politicians performances in Brazil. The analysis of who these political actors are and how they influence national politics is a matter of great relevance in the Political and Social Sciences today. The understanding of how government managed elections data, how political campaign funding data are made available and how society's appropriation of this data occurs are issues to which the Information Science methodological and theoretical apparatus can contribute.

Considering the recent years have been the longest period in Brazil in which a democratic regime has been able to operate in a relatively stable manner,¹ researches that analyse the drivers that strengthen or threaten democracy are welcome on Social and Political Sciences. One of the main aspects that influenced the country’s political scene was the growth of evangelical influence groups, the so-called “Bancada Evangélica” (Evangelical Parliamentary Front). These “Bancada” imposed a conservative agenda that seeks to contest the secularity of state and restrict minority rights and civil liberties (Carranza and Cunha 2018) and, on the last presidential campaign, they actively supported the election of the current far-right president Jair Bolsonaro. This influence is not an isolated phenomenon in Brazil. The presence of evangelicals in politics has been taking place for more than three decades in all of Latin America (Freston 2008).

¹ Since the beginning of the Republic in 1891, which was proclaimed and governed by the military in its first 8 years, until the constitution of 1988, Brazil went through brief democratic periods interspersed with two dictatorial regimes: the Dictatorship of the “Estado Novo” when Getulio Vargas ruled from 1930 to 1946 and the Civil-Military Dictatorship from 1964 to 1985.
In fact, the generic term “evangelical” does not reflect the multiplicity of social, economic, religious and moral groups that fit within this religious segment. However, in Brazil, according to Mafra, “[…] given the public visibility this segment has gained in public opinion, a certain consensus has been forged by referendum on the term ‘evangelical’ as a comprehensive category” (Mafra 2001, 7). It is worth noting that the generic term “evangelical” used in Brazil is the same as “Protestant” in the USA and Europe, while “Protestant” in Brazil is similar to “Main Lane Protestant Church” in the USA (Mafra 2001).

As the Brazilian Census is one of the most used scientific classification instrument of religion studies in Brazil, a comparative review of how it classifies religions can bring relevant contributions to the studies over the processes of constructing knowledge organization systems of religion, also to theoretical and empirical studies of religion and its relation to politics. Effectively, who are those Christian groups that appear on the political scene? Do they represent the population contingent that calls itself evangelical? Are they homogeneous or is there a prominent protestant denomination that leads this movement? To start to answer those questions, the first step is to identify how part of the recommended literature and the bibliographic schemes classify Religion and then how Christian denominations are classified in relevant classifying schemes in the world. This paper details how this research is starting to work with these questions, in an introductory exploratory overview.

2.0 The method

This study is an exploratory, qualitative, theoretical and documental research. It aims to make a comparative review of the classification of religion used in the Brazilian IBGE Census of 2010\(^2\), with a specific attention on the protestant groups and denominations. It compares the traditional bibliographic classification schemes on Dewey Decimal Classification (DDC), Universal Decimal Classification (UDC) and Colon Classification (CC), showing their similitudes and differences with the classification constructed by the Brazilian Census.

The bibliographic research explored specially three sources: the classifications 1) of religions in different and influent bibliographic schemes; 2) of the Protestants on specialized literature; 3) of the Protestants in the Brazilian census.

3.0 Religion Classification

“Omnis determinatio est negatio” (Spinoza cited in Lenin 2011, 111). Every determination excludes, in purely logical terms. In ethical and political terms, what means each consecrated set of determinations, as the main bibliographic classifying schemes? Which contending worldviews operate in each case, for what reasons, with what consequences?

Classifying religions is an arduous task. There is no consensus on the best way to classify them. In fact, there is not even consensus on what differs religion, denomination, sect or cult (Liebman, Sutton, and Wuthnow 1988). The most widely used CS in the West stem from European culture, created at the height of Imperialism Era, before the

\(^2\) IBGE. 2011. Censo Demográfico 2010: Características Gerais da População, Religião e pessoas com deficiência
Second World War. Christian religions were the main ones in the reality in which Dewey and Otlet lived, while Asia, Africa and Latin America (the now called Global South) were exclusively suppliers of raw material to Europe and USA. That hegemonic culture was reflected on both CS: DDC and UDC just constructed the Religion class (named 200 on both schemas) reserving the majority of sub-classes dedicated to Christianity and reserves the “Others” (290) to all other religions, putting together ancient religions (Greek and Roman Mythology) in the same level of Islam, Hinduism or Buddhism. Indeed, the UDC was more inclusive than the DDC because it specified much more religions on its main classification schema. Nevertheless, it kept the same structure giving most of the main classes dedicated to Christianity.

It may not be fair to blame the librarians and professionals of that time who participated in the making of the classification schemes. Pragmatically, how many books related to Hinduism, Islam or the religions of China or Japan arrived on the Western countries? Would it be rational to classify religions according to their importance if the number of documents arriving in the libraries (and to the classifiers) was minimal? For instance, in UDC there was a whole classification section to Judaism (with 49 sub-classes) while other religions with many more believers at that time did not have such relevance. Was there a political or socio-economic influence to include so many sub-classes in the Judaism scheme? Or were there simply more documents about Judaism in Western libraries, since the Jewish presence in the West was relevant at the beginning of the 20th century and because Christianity itself came from Judaism, maintaining with Judaism an ambiguous but intense relationship throughout the centuries?

These schemas based mainly on Christian classes changed when Ranganathan built the Colon Classification (CC). That classification, for the first time, represents eastern religions in detail. Apparently, the classification proposed by the CC listed the religions in a historical order, i.e., ordering the oldest religions in the first positions of the classification scheme. However, it was not possible to find a bibliographic reference that would make explicit the choice for this methodological option. As can be seen in Table 1, CC was very concise classifying the Christian religions. One can note that, in the same way that the CDD and the CDU did not perform a greater detailing for the other Eastern religions, the CC did the same for Christianity. In the comparative table, we deliberately emphasize the terms sects, movements, denominations and churches in order to demonstrate how such terms appear without apparent criteria.

We also analysed how the Chinese system performed the classification on religions. Although its use is restricted to China, it has great relevance since China has one sixth of the world population. Still in the 20th century, just after the Communist Revolution in China, four different classification schemes were created, each for different purposes. The four systems had, however, a common basis quite distinct from the Western Aristotelian vision. While Western CS work philosophy as their first hierarchical item, the Chinese classification brings Marxism as its first class and the class of religion appears as a mere appendix to the class of Philosophy. In the 1970s, there was an effort to create a single CS carried out by the University of Beijin, but this new system still keeps the same ideological bases as the four initial systems (Studwell, Wu, and Wang 1994). Unfortunately, during this literature review, it was not possible to verify deeply the classification of religions within the Chinese system, but it was interesting to verify the lack of relevance of the subject.
Indeed, several studies addressed how biased the western classification schemas are when the subject is religion. Vanda Broughton identified three main areas where bias can occur: “an illogical order, or distribution of notation, that causes one system to appear as dominant, use of vocabulary that has a strong flavour of one system or is special to that system and inadequate provision of detail other than for the 'favoured' religion” (Broughton 2000). The author notes, however, that it is possible to minimize the bias through the use of facet analytical techniques and that the most recent UDC revision of Class 2 can be compounded and can achieve a better degree of specificity. Idrees and Khalid (2009) wrote a study on the Islam classification and proposed amendments and expansion in order to give a better guidance to LIS professionals.

Finally, the article by McIlwaine and Mitchell (2006) first at ISKO Conference and then replicated in one of the UDC’s own reports, suggested ways to minimize the impact of classification bias through an auxiliary table listing religions in chronological order, i.e., from the oldest to the newest religions. In the same article, they agreed with Broughton that the revision of UDC class 2 would bring benefits to such classification system.

**Table 1: Christian Religions Equivalence between CDD, CDU and CC**

<table>
<thead>
<tr>
<th>CDD</th>
<th>CDU</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>281 Early <strong>church</strong> and Eastern <strong>churches</strong></td>
<td>281 Primitive Churches. Eastern Churches.</td>
<td>61. Early Churches</td>
</tr>
<tr>
<td></td>
<td>281.93 Russian Orthodox Church</td>
<td>618. Russian</td>
</tr>
<tr>
<td></td>
<td>281.94 Greek Orthodox Church</td>
<td>611. Greek</td>
</tr>
<tr>
<td></td>
<td></td>
<td>613. Armenian</td>
</tr>
<tr>
<td>282 Roman Catholic <strong>Church</strong></td>
<td>282 Roman Catholic Church</td>
<td>62. Roman Catholic</td>
</tr>
<tr>
<td>283 Anglican <strong>churches</strong></td>
<td>283 Episcopal Churches – not Roman Catholic (Protestants)</td>
<td>63. Protestant</td>
</tr>
<tr>
<td></td>
<td>283(410.1) Anglican Church</td>
<td></td>
</tr>
<tr>
<td></td>
<td>283(73) Episcopal Church on USA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>283.5 Old Catholics (?)</td>
<td></td>
</tr>
<tr>
<td>284 Protestant <strong>denominations</strong> of Continental original</td>
<td>284 Continental Protestant Sects</td>
<td></td>
</tr>
<tr>
<td></td>
<td>284.1 Luterans</td>
<td></td>
</tr>
<tr>
<td></td>
<td>284.2 Calvinists</td>
<td></td>
</tr>
<tr>
<td></td>
<td>284.3 Utraquistas. Tabortas.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>284.4 Coterões. Gazaristas.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>284.5 Huguenotes. French Movements</td>
<td></td>
</tr>
<tr>
<td></td>
<td>284.6 Moravian Brothers. Hernutos.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>284.98 Old Lutherans. Free Lutheran church</td>
<td></td>
</tr>
<tr>
<td>285 Presbyterian <strong>churches</strong>, Reformed centered in America</td>
<td>285 Puritanism</td>
<td>65. Puritanism</td>
</tr>
<tr>
<td></td>
<td>285.1 Presbyterian Churches</td>
<td>64. Presbyterian</td>
</tr>
<tr>
<td></td>
<td>285.8 Brownists. Barrowists.</td>
<td></td>
</tr>
</tbody>
</table>

---

3 The CDU that we accessed was in Portuguese. We translated the terms to English. But we could not find the English word of some of them (Coterões, Hernutos, Utraquistas and Tabortas) for them.
4.0 Protestant Classification Issues

People who profess the Christian faith and are part of religious groups or churches that emerged from the Protestant reform initiated by Luther in the 16th century are generally called Protestants or evangelical. However, how to categorize a sect or church created in the 21st century? From the same theological dogmas of the original groups or just from the self-denomination of its representatives?

For example, according to Pew Report “Global Christianity” (2010), there were 801 million Protestants in the world and they represented 37% of the global Christian population. The report categorized Christianity in four major groups: Orthodox, Catholic, Protestants and “Other Christians”. The latter group refers to all Christian religions that do not fit into the first three. The classification of Catholics and Orthodox seems to be easy to accomplish, as they are churches that have well-defined theological, historical and geographical attributes over almost 18 centuries. However, how to make the difference between Protestant and other types of Christians, that are quite emblematic as Mormons, but who make a stand of not being classified as Protestants?

Such difficulty in categorizing the Protestants was noted from the beginning of the study of Protestantism. One of the most respected bibliographies on the subject was the Protestant Ethics and the Spirit of Capitalism, by the German sociologist Max Weber. In this work, Weber tried to catalogue the various groups of Protestants that existed in his time and already indicated the difficulty in performing such categorization:

“we can only do this by presenting religious ideas with the logical consistency of an "ideal type", which is only rarely found in historical reality. Precisely because of the impossibility of drawing clear boundaries in historical reality, our only hope in researching the most coherent of its forms is to tune in with its more specific effects.” (Weber 1999, 90).

The author categorized the Protestants into five major groups (Lutherans, Calvinists, Pietists, Methodists and Anabaptists), based on the moral dogmas that each group followed. The interesting categorization adopted by Weber came from the need to perceive which groups fit more in the so-called spirit of capitalism, the new ethics that had emerged in the same historical period as the Protestant reform and that became stronger precisely in the countries that embraced capitalism with more intensity. In other words,
the methodological option of categorization had the objective of understanding how the new capitalist ethos was strengthened in each group.⁴

Bringing the discussion to present time, the Pew Institute uses two kinds of categorization: by movement and by denomination. The first one categorizes Christians in three major movements: Pentecostals, Charismatics and Evangelical and the criterion used for this categorization is based on the central dogmas of each religious group, independent of the historical and geographical context. The second one, by denomination, categorizes them by their history and origin. This last CS is the only one that resembles the bibliographic classification.

In Brazil, the term Protestant is used usually on the academic and specialized contexts. As we noted on the introduction, non-Catholic Christian are generally called “evangélicos”. Mainstream media and everyday talking usually refer to them as “crentes” (believers), in a pejorative connotation. However, this so-called evangelical population is heterogeneous from a theological, economic and political point of view.

Historically, the arrival of the first evangelicals in Brazil coincided with the opening of the ports with the coming of D. João and the permission for other religions than the Catholic to settle in the Portuguese colony. However, this first wave of Protestants had only the objective of serving the foreigners who arrived together with the Portuguese court and the act of evangelization was very timid. It was only in the 1910s that the first missionaries arrived in Brazil with the intention of evangelizing the population. The Assembly of God and the Christian Congregation were founded at that time (Mariano 2004). The second significant wave of protestants began to take root in the 1950s and 1960s with the founding of the “Igreja do Evangelho Quadrangular” (Foursquare Gospel Church) and then the “Brasil para Cristo” (Brazil for Christ), “Deus é Amor” (God is Love) and “Casa da Bênção” (House of Blessing) churches. Specifically, the Foursquare Gospel Church began to act with greater force, and as Mariano attests, “besides the emphasis on healing, this Pentecostal aspect was noted for the intense use of the radio and the itinerant preaching with the use of canvas tents” (2004, 123).

From the end of the 1970s on, the number of evangelicals started to grow expressively with the emergence of the so-called Neo Pentecostal churches. Originating from Methodism, this new aspect of Protestantism brought a new form of philosophy: the so-called Prosperity Theology, which moved away from the asceticism of the Methodists and brought a new form of belief that valued material goods. In general terms, it defines that the greater man's faith, the greater his material prosperity.

Unlike the traditional churches that had their origins in European countries and the USA, this new side of Protestantism grew out of churches shaped by Brazilian pastors and some of these churches became corporations with enormous financial, media and political capacity. The most common nomination of them is Neo-Pentecostals. Among these churches, we should highlight the “Igreja Universal do Reino de Deus (IURD)” (Universal Church of the Kingdom of God), the “Graça de Deus Internacional” (Grace of God International), Sara Nossa Terra (Heal Our Earth) and “Renascem Cristo”

⁴ Keeping in mind this Weberian methodological option, we also ask about the relation between nowadays capitalist’ ethos and the many protestant groups in Brazil, from a critical study of the political performances of the politicians that are supposed to represent them. A complementary methodological approach, so, is to relate ethos and ideology, in the sense of false consciousness. This is to say, morality as a lure to obtain votes. We will come back to this point sooner.
(Reborn in Christ). Still according to Mariano, in the 80's alone the IURD grew 2,600% and at the end of the 90's it was estimated that it had more than 2 million followers (2004, 125). The organization controls the 2nd largest TV station in Brazil, hundreds of radio concessions and since then it has expanded internationally: in 2001 it was already present in 80 countries, having more than 1000 temples (Oro 2004).

5.0 The IBGE Census

The Brazilian Institute of Geography and Statistics (IBGE) is a public institute of the Brazilian federal administration and the main provider of country’s demographic data. As stated on their site, “such information meets the demands of several types of segments of civil society, as well as the bodies at the federal, state and municipal level”. And one of their main missions is to conduct the National Census which occurs every 10 years.

As far the research goes, just 6 countries (Brazil, Canada, México, Peru, Jamaica and Haiti) on South, Central and North America conducted a religion inquiring during their National Census. All other countries on the continent rely on private institutes to estimate how many practitioners exists and what are their religion. The research had focus just on this 3 sub-continents because it will be very time consuming to do it to all countries of the world. An interesting case occurs on the United States where there is a specific law that prohibits the National Census to conduct any inquiry about religion. The law promulgated in 1950 came at the same time as the horrors of Nazism were discovered after the end of the second world war, and the concern that the State might distinguish its citizens on the basis of their religion may have been the cause of the implementation of the law. However, during the research, no evidence was found that the motivation for such law occurred because of the Jewish Holocaust.

Since 1872, the Brazilian census had included a question on Religion and until 1970 the Roman Catholic reigned absolute with 91.8% of the population declared themselves Catholic. However, after the 1991 Census, there were significant changes in the religious composition for the first time when there was a growth in the number of respondents who declared themselves evangelical. In the 1980s, the percentage of the population segment that declared itself evangelical was only 6.6% and rose to 9.0% in the following decade, and up to the present time the share of the evangelical population continues to grow steadily: according to the 2010 census, the Brazilian total population was made up of 22.2% evangelicals. There was also a slight growth of those who have declared themselves to be spiritists and of all other religions but the most notable fact is that the Catholic population have been diminished in a regular pace. Due to this new scenario, IBGE has sought a more appropriate categorization of existing religions, relying on the advice of a Brazilian NGO specialized in Religion called ISER (Institute of Religious Studies).5

Founded during the 1970s, ISER has a tradition not only in the area of religion but also in the defence of human rights, public security and the environmental issues. It regularly publishes two renowned peer reviewed journals in the area and books about religion and human rights, in addition to promoting ecumenical meetings among the various religious representatives. The schema they elaborated starts with three main classes: Traditional Protestants, Pentecostals and Others. The traditional protestants are

5 http://www.iser.org.br/site/o-iser-ingleshe/
those from European and North American origin, the Pentecostals are those created by Brazilian pastors and “Others” are those who don’t fit on the first two categories.

The application of the categorization was much criticized at the time. The methodological choice of ISER was concerned with naming the churches but the census technician could not identify the church in a proper way. As the census form came with an open question: "What is your religion?" without offering a list of options, the critics were in doubt as to how the interviewee would have answered and how precise the technician who wrote down the answers would reproduce what was said (Camurça 2014). Given the controversy, ISER asked IBGE for the database with the individual responses of the respondents (even if anonymized) so that it would be possible to assess how accurate the collection was. However, IBGE denied the request.

The final categorization used and numbers collected by the Census was summarized in Table 2. In this table it is possible to see that the name of the Pentecostal churches appears differentiated from the religions categorized as Churches of Mission. The latter are the churches that appear in both the CDD and the CDU. It is worth noting the significant number of the “Evangelical undetermined” category, which represents more than 20% of all non-Catholic Christians. If we add up the number of ‘Evangelicals of Pentecostal origin – others’, the percentage rises to 33% of the total number of Evangelicals. That is, this number confirms a great pluralism among the evangelicals. One possibility was that many faithful adherents of stigmatized groups chose not to declare their attachment to a specific church, even more because during the census period there were many corruption scandals linked with evangelical politicians and struggles in the evangelical milieu (Sottani 2010).

Table 2: Protestant churches according to IBGE Census 2010

<table>
<thead>
<tr>
<th>Churches of Mission</th>
<th>Population</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luteran Church</td>
<td>999,498</td>
<td>2.00</td>
</tr>
<tr>
<td>Presbiterian Church</td>
<td>921,209</td>
<td>1.84</td>
</tr>
<tr>
<td>Metodist Church</td>
<td>340,938</td>
<td>0.68</td>
</tr>
<tr>
<td>Baptist Church</td>
<td>3,723,853</td>
<td>7.45</td>
</tr>
<tr>
<td>Congregational Church</td>
<td>109,591</td>
<td>0.22</td>
</tr>
<tr>
<td>Adventist Church</td>
<td>1,561,071</td>
<td>3.12</td>
</tr>
<tr>
<td>Other Mission Evangelical</td>
<td>30,666</td>
<td>0.06</td>
</tr>
<tr>
<td><strong>Churches of Mission Sub Total</strong></td>
<td><strong>7,686,826</strong></td>
<td><strong>15.39</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Evangelical Pentecostal</th>
<th>Population</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Igreja Assembléia de Deus (Assembly of God)</td>
<td>12,314,410</td>
<td>24.65</td>
</tr>
<tr>
<td>Igreja Congregação Cristã do Brasil</td>
<td>2,289,634</td>
<td>4.58</td>
</tr>
<tr>
<td>Igreja o Brasil para Cristo</td>
<td>196,665</td>
<td>0.39</td>
</tr>
<tr>
<td>Igreja Evangelho Quadrangular</td>
<td>1,808,389</td>
<td>3.62</td>
</tr>
<tr>
<td>Igreja Universal do Reino de Deus</td>
<td>1,873,243</td>
<td>3.75</td>
</tr>
<tr>
<td>Igreja Casa da Benção</td>
<td>125,550</td>
<td>0.25</td>
</tr>
<tr>
<td>Igreja Deus é Amor</td>
<td>845,383</td>
<td>1.69</td>
</tr>
<tr>
<td>Igreja Maranata</td>
<td>356,021</td>
<td>0.71</td>
</tr>
<tr>
<td>Igreja Nova Vida</td>
<td>90,568</td>
<td>0.18</td>
</tr>
<tr>
<td>Evangélica renovada undetermined</td>
<td>23,461</td>
<td>0.05</td>
</tr>
<tr>
<td>Evangelical Community</td>
<td>180,130</td>
<td>0.36</td>
</tr>
</tbody>
</table>
Other Evangelic Pentecostals | 5,267,029 | 10.54
Undetermined Evangelical | 9,218,129 | 18.45
Not identified as Evangelic or Catholic by IBGE
Other Christian religiosities | 1,461,495 | 2.93
Church of Jesus Christ of Latter-day Saints | 226,509 | 0.45
Jehovah's Witnesses | 1,393,208 | 2.79
Grand Total | 49,962,264 | 100

6.0 Conclusion

In Brazil, a popular proverb states that in the dinner table, you should not talk about religion, politics and soccer. However, in the last 35 years, two of those controversial topics were melded on the country’s incipient democracy. The turmoil that promoted a presidential impeachment under suspicious allegations and jail former president Lula da Silva had direct participation of evangelical actors. To understand more precisely who they are, what groups finance them and how democracy can protect itself from those groups are vital to protect civil liberties and rights. It is vital to avoid the persecution of minorities, or even majorities, “deviant” from the set of conservative values rhetorically or effectively defended by “evangelicals”. These persecutions take place in Brazil against religious minorities, specially of African origins and indigenous people, to non-heteronormative people, feminists, the arts in general, mass culture, critical social sciences and even modern science, insofar as it contradicts more literal readings of the Bible – creationism x evolutionism, for example. These persecutions take the concrete form of projects of law, censorship of plays and art expositions, public schools curricula changes (excluding the obligation of the study of African history, philosophy and sociology), targeting of publicity funds for more friendly media corporations, preaching for millions in churches, radio and television stations, owned or rented by these churches etc. (Almeida 2019).

Besides, the conservative moralist discourse of protestant politicians mobilize the affections of the “evangelicals” in support of political groups whose actions in the field of economics have a clear neoliberal bias. These discourses were often widespread defunded out of Brazil in the form of bizarre fake-news against political opponents, through social networks of believers (illegally, during the electoral period)6. The main point here is that behind moralism, the flexibilization of the labour legislation, favouring the employers, the scrapping of public services, from health to education and even water supplies, actually contradicts the concrete interests of these same believers, mostly workers. In other words, moralism acts as a kind of Trojan Horse, camouflaging the interests that are effectively at stake.

This paper maps the construction of protestant classification on several knowledge organization systems in order to understand how social researchers can critically use those classifications on their scientific researches. By comparing those classification with the IBGE Brazilian census, we could perceive how the census classification is adherent to the classifications used in worldwide relevant bibliographic systems.

---

6 We must note that in Brazil, broadcasting operations, even private ones, are public concessions. By the law, they should be committed with the promotion of accurate information, “serious” and popular culture, tolerance.
As we see, there are many possibilities to scrutinize the classification of evangelical groups and research in this subject can better clarify how these population moves among the various religious denominations in the evangelical field. On our ongoing research, we could identify how the census of other countries outside the American continent deal with Religion inquires and how the knowledge organization systems adhere to each other. Studying different CSs over a controversial theme as religion, in many countries and cultures, can open space to new and less ethnocentric biased intercultural and multifaceted classification schemes, what Otlet and Ranganathan aimed.

References
Cacao, Cocoa, or Cocoa? 
Reconciliation of Taxonomic Names in Biodiversity Heritage Library

Abstract:
The Biodiversity Heritage Library (BHL) currently hosts more than 150 thousand titles, and 57 million OCR-scanned pages on biodiversity literature dating back to the 16th century. While great research efforts have been conducted to extract taxonomic names in BHL’s literature the issue of name reconciliation has yet to be studied. Through the use case of Theobroma cacao, commonly known as chocolate plants, this research aims at presenting a framework to reconcile species names in BHL by merging external taxonomies. We demonstrate this by using a logic-based, taxonomy alignment approach to match variations of species and subspecies names of Theobroma cacao from four major biodiversity sources: the Encyclopedia of Life (EoL), Integrated Taxonomic Information System (ITIS), Global Biodiversity Information Facility (GBIF), and the United States Department of Agriculture PLANTS Database (USDA Plants).

1.0 Introduction

Consider the following hypothetical scenario: Charlie is a rising researcher in plant biodiversity with an interest in economic plants and wants to find more descriptions about such species and their family trees from historical texts. Determined to use the canonical online resource for biodiversity texts – the Biodiversity Heritage Library – Charlie starts with a quick search with cocoa, the ingredient in chocolate bars. There are two seemingly unrelated species names. She then turns to GBIF, another popular online resource for biodiversity information, to perform further searches. She found that in GBIF there are two names listed: Theobroma cocao and Theobroma cacao. Unsure which names are the approved scientific name, she tries both and finds 15 other variations of scientific names for Theobroma cacao shown in BHL. These 15 names are listed alphabetically, but without further information of how the terms relate to each other.

With the exponential growth of biodiversity information and data, it has become increasingly difficult to reliably retrieve species based on species names, as shown in the example above. The Biodiversity Heritage Library (BHL) houses OCR-scanned pages of legacy biodiversity literature from natural history museum collections and other partnering institutions. This literature dates back to the 16th century, but oftentimes key taxonomic papers are obscured by the evolving, different variations of species names. While efforts have been put in leveraging natural language processing (NLP) and text mining methods to extract scientific names from the texts in BHL, such as developing a series of text mining tools (Page 2011, 2013), discussing taxonomic name recognition from texts (Wei, Heidorn, and Freeland 2010), or automating the process of extraction of names from texts (Batista-Navarro et al., 2015), little has been done in organizing and reconciling these scientific names to provide a better species name representation.
Through the use case of the species *Theobroma cacao*, commonly known as chocolate plants, our research aims to develop a framework for reconciling species names in BHL by merging external taxonomies. We propose to employ a logic-based, taxonomy alignment approach (Franz et al., 2015, 2016; Cheng and Ludäscher, 2019) to match variations of species and subspecies names of *Theobroma cacao* from four major biodiversity sources: the Encyclopedia of Life (EoL), Integrated Taxonomic Information System (ITIS), Global Biodiversity Information Facility (GBIF), and the United States Department of Agriculture PLANTS Database (USDA Plants).

### 2.0 Related Work

#### 2.1 Quality issues in aggregated databases such as GBIF, BHL

Data quality issues in aggregated biodiversity databases are not uncommon. Numerous studies have discussed the issues as well some practical solutions towards these aggregated databases. Parr et al. (2012) discuss how the rapid growth of biodiversity information from various repositories exacerbates the discovery and integration of the knowledge to the “Tree of Life”, and call for a linked data solution to connect all different aggregated databases. Franz and Sterner (2018) point out how these aggregators underplay a “taxonomic backbone” in their design systems, and shift responsibilities of data quality issues to the data source providers. The authors conclude that rather than correcting datasets from the root providers, providing services that enhance the taxonomic concepts in these systems would increase the trust and collaboration among systematists and the aggregator communities. While the discussions on data quality mostly circulate around aggregators such as GBIF, extracting scientific articles and finding species names in BHL has also been proven to be an unresolved concern.

#### 2.2 Extracting names in BHL

BHL uses Global Names Recognition and Discovery (GNRD) services to analyze the OCR-scanned texts and identify any string that can potentially be scientific names. Many studies have proposed using natural language processing (NLP) and text mining methods to enhance the recognition and extraction of scientific names from BHL. For instance, NLP techniques have been developed to support the extraction of taxonomic names and morphological character from taxonomic descriptions (Thessen, Cui, and Mozzherin 2012). Page (2011, 2013) has developed a series of text mining tools (BioNames, BioTor) that aims to enhance the extraction of names and retrieval ability of BHL. Recently, Page (2019) proposed an approach to link taxonomic names from other databases to papers from BHL with the attempt to leverage existing resources. These prior works show great potential in enhancing the extraction and recognition of names from BHL on top of its native embedded service on the GNRD API. However, how to best utilize, organize, and represent the extracted information from BHL warrants further research.

![Figure 1. BHL’s current search results only return a flat structure of species names](image-url)
2.3 Reconciliation of names

Name recognition and extraction are of growing interest within the context of BHL to provide semantics for unstructured data (Wei, Heidorn, and Freeland 2010); nevertheless, the extraction of names is only the first step towards retrieving species information from text, the reconciliation of species names is the next unsettled territory to many biodiversity experts. In Franz et al. (2016), the authors reconciled 11 different taxonomies across 126 years of time span of the Andropogon complex and concluded that the meanings of scientific names can change significantly over time. Analogously, extant literature in Knowledge Organization also emphasized how names are a diachronic concern by both biologists and information scientists alike (Blake 2011). Our prior work discussed the reconciliation of names in evolving, disputed geo-entities (Cheng and Ludäscher 2019).

To date, BHL remains a cornucopia of historical taxonomic names and species information. However, the current organization of species taxonomies in BHL makes it hard to relate species names and determine the credibility of species hierarchies. This paper proposes to reconcile species names from external taxonomies and link this value-added information with information from BHL to obtain an even more informative and user-friendly presentation of search results.

3.0 Use Case: Theobroma cacao

We started with a single species to investigate the current taxonomic backbone of BHL. The species we chose to examine is commonly known as cacao (or cocoa) tree, scientifically known as *Theobroma cacao*. Upon searching *Theobroma cacao*, BHL performed a full-text search and returned about 4,000 search results on all the publications and their metadata (title, author, date, page number) relating to the search term. 15 scientific names were listed alphabetically on the BHL interface. However, the relationship of how each term relates to *Theobroma cacao*, whether hierarchical, synonyms, or siblings, was unknown to the users. For instance, when clicking a term among the 15 names such as *Cacao theobroma*, BHL returns a list of bibliography that contains the keyword. The current BHL structure is shown in Figure 1. Because information about where species and higher taxa occur in a taxonomic hierarchy is essential for taxonomists, biodiversity experts, and researchers, we are particularly interested in how BHL presents the information of scientific names, and whether a better approach to organize these names can be provided.

Figure 2. Our proposed taxonomic structure framework.
3.1 Method
We propose a framework to gather species taxonomies through external databases. Specifically, we compare and merge taxonomies of species in four major sources: EoL, GBIF, ITIS, and USDA Plants (Figure 2). In this study, we specifically investigate the inclusion of subspecies into the taxonomic backbone of BHL.

3.2 Data Collection
Our data was collected in December, 2019. Detailed descriptions of the data are stated below. The taxonomies of the four different external databases are not mutually exclusive and may overlap with one another.

3.2.1 Encyclopedia of Life (EoL)
EoL is one of the aggregated databases that annotates preferred scientific names, common names, and synonyms of species. Notably, EoL maintains curated dynamic hierarchies for each species. EoL staffs manually curate and edit the taxonomic information as needed when suggested by biodiversity experts and likewise communities.

3.2.2 Global Biodiversity Information Facility (GBIF)
GBIF is one of the most popular sources for biodiversity information, including species occurrences, publication, peer-reviewed data, and etc. GBIF published a backbone taxonomy containing taxonomies from EoL, IUCN Red lists, published papers, and more. The taxonomic information is GBIF is updated via an automated process, and Catalogue of Life (CoL) is the primary source that GBIF compares the taxonomies upon.

3.2.3 Integrated Taxonomic Information System (ITIS)
The goal of ITIS is mainly to provide species names and taxonomic information. It contains approved species taxonomy by biodiversity experts and its Taxonomy Working Group (TWG). Performing a search on any species will result in a list of accepted or not accepted species names, followed by species hierarchies, expert references, or other sources.

3.2.4 United States Department of Agriculture PLANTS Database (USDA Plants)
The scope of USDA Plants is slightly narrower than the previous mentioned counterparts. Performing a search on species names will lead users to the species classification, and subordinate taxa which documented species parents (genus, family), and subspecies.

Table 1 shows the subspecies we collected from the four external sources. We considered all the infraspecific epithets (names) as children of a species, so species names that contain keywords such as subsp., ssp., infrasubsp., f., fm., var., were all grouped as subspecies in our analysis. In particular, we viewed subsp. equivalent to ssp. (subspecies), and f. equivalent to fm. (form). That said, if a database consists of both Theobroma cacao subsp. cacao and Theobroma cacao ssp. cacao, we only keep one name for the purpose of analysis.

---

1 EoL information page: https://eol.org/docs/what-is-eol/whats-new
Table 1. A list of subspecies in the four sources examined. *Data are collected December 2019. Prefix Theobroma cacao omitted; changed punctuations (.) to underscores (_).

<table>
<thead>
<tr>
<th>EoL</th>
<th>GBIF</th>
<th>ITIS</th>
<th>USDA Plants</th>
</tr>
</thead>
<tbody>
<tr>
<td>ssp_cacao</td>
<td>subsp.paeano</td>
<td>ssp.cacao</td>
<td></td>
</tr>
<tr>
<td>ssp.sphaeroarpa_A_Chevalier_Cuatre</td>
<td></td>
<td>ssp.sphaeroarpa_Chevalier_Cuatre</td>
<td></td>
</tr>
<tr>
<td>f_lacandonesi_Cuatre</td>
<td></td>
<td>f_lacandonesi_Cuatre</td>
<td></td>
</tr>
<tr>
<td>f_pentagonon_Bernoulli_Cuatre</td>
<td></td>
<td>f_pentagonon_Bernoulli_Cuatre</td>
<td></td>
</tr>
<tr>
<td>f_leiocarpum_Bernoulli_Ducke</td>
<td></td>
<td>f_leiocarpum_Bernoulli_Ducke</td>
<td></td>
</tr>
<tr>
<td>ssp_leiocarpum_Bernoulli_Cuatre</td>
<td></td>
<td>other_infraespesc</td>
<td></td>
</tr>
</tbody>
</table>

3.3 Logic-based Taxonomy alignment approach

3.3.1 Taxonomy

We define taxonomy $T$ as a hierarchical, tree structure of terms (or names): Each node in the taxonomy has only one parent (with the exception of the root node having no parent). Sibling nodes typically represent disjoint taxa, i.e., two nodes on the same level in the tree are considered mutually exclusive. When considering the children of a node, we could either assume that all children are known (i.e., we know or assume that no other children nodes exist), or that there might yet be unknown children. In case of the latter, we introduce a placeholder node called “other to design for future possible changes (e.g. *Theobroma_cacao_other*).

3.3.2 Taxonomy Alignment Problem (TAP)

To compare two taxonomies $T_1$ and $T_2$, we first identify a set of articulations (relations) used to describe how concept X in $T_1$ relates to concept Y in $T_2$. The region connection calculus RCC-5 (Randell, Cui, and Cohn 1992; Cohn and Renz 2008) can be used to define specific articulations: equals, overlaps, disjoint, includes, is_included_in. Then, we input these constraints to Euler/X (an Answer Set Programming, Python-based tool), which provides us with different merged solutions for these pairwise comparisons. Euler/X will either conclude with (1) an inconsistent outcome with zero Possible World (PW) ($n=0$); (2) a single, uniquely merged PW $T_3$ ($n=1$), usually the desired outcome; or (3) multiple merged PWs $T_3$ ($n≥2$), where each world is a possible reconciliation of how the two taxonomies can be merged. A simple TAP is shown in Figure 3, where a few relations are marked as equivalent ‘=’, and the ‘other’ nodes are left with open relations. The resulting PW shows that all the children nodes in both taxonomies are equivalent (in grey boxes). Details of the Euler/X tool workflow and descriptions for implementation are explained in Cheng et al.(2017).

4.0 Results

From Table 1, it shows that some of the subspecies names are equivalent, but with variations on choosing either using *ssp.* or *subsp.* to represent subspecies, using *f.* or *fm.*
to represent forms, and using different authorship spellings (e.g. A. Chevalier vs. A. Chev, L. vs. Linnaeus, etc). For such cases, we regard the two names as equivalent. Table 1 shows that ITIS and USDA plants provide fewer subspecies names. Therefore, we begin our alignment with these two taxonomies that are less complex than others.

Given that one of the limitations of our taxonomy alignment approach is that only two taxonomies can be compared at once, we executed six pairwise alignments in total to compare each taxonomy with the other three ($T_{ITIS}$-$T_{USDA}$, $T_{EoL}$-$T_{ITIS}$, $T_{ITIS}$-$T_{USDA}$, $T_{GIBF}$-$T_{ITIS}$, $T_{GIBF}$-$T_{USDA}$, $T_{EoL}$-$T_{GIBF}$).

4.1 $T_{ITIS}$-$T_{USDA}$

We begin our alignment with these two taxonomies. Evidently, as shown in the PW of Figure 4, every node is considered congruent (in grey round boxes), meaning the two names from each taxonomy is equivalent. This means instead of six pairwise alignments of the four taxonomies, we can reduce our number of alignments to four given the merged Possible World (PW) of $T_{EoL}$-$T_{ITIS}$ and $T_{EoL}$-$T_{USDA}$, $T_{GIBF}$-$T_{ITIS}$ and $T_{GIBF}$-$T_{USDA}$ will be in exactly the same structure.

4.2 $T_{EoL}$-$T_{ITIS}$, $T_{EoL}$-$T_{USDA}$

Considering that $T_{ITIS}$ and $T_{USDA}$ are equivalent, here we only show the result for $T_{EoL}$-$T_{ITIS}$ (Figure 5). $T_{EoL}$ provides more subspecies information than either $T_{ITIS}$ or $T_{USDA}$. As a non-biodiversity expert, we cannot assert how these subspecies relate to each other, therefore, we left the relations open in the input alignments. As a result, all the extra subspecies in $T_{EoL}$ are inferred to be also merged under “TCOther” in $T_{ITIS}$ (and $T_{USDA}$) (Figure 5, PW). This indicates that $T_{EoL}$ may be a bigger taxonomy than the other two and $T_{EoL}$.Theobroma cacao includes both $T_{ITIS}$.Theobroma cacao and $T_{USDA}$. Theobroma cacao.

Figure 4. Input alignment (top) and output Possible World (bottom) for $T_{ITIS}$ and $T_{USDA}$

Figure 5. Input alignment (top) and output Possible World (bottom) for $T_{EoL}$ and $T_{ITIS}$
4.3 TGBIF and TITIS, TGBIF -TUSDA

Similar to the result in 4.0.2, since there were no direct counterparts for many of TGBIF’s subspecies, we left the nodes' relations open without linking them to anything in TITIS or TUSDA. The merged PW is also indicative that TGBIF is more granular in terms of subspecies, and the subspecies should be merged under TITIS’s “Other” category. (See Appendix for visualizations of 4.3 result)

4.4 TEdl and TGBIF

Figure 6 shows the two merged PWs of TEdl and TGBIF. Notably, for this pair of taxonomies, we ended up with more than one PWs, potentially due to the influence of the “other” category. Specifically, the differences between the two PWs are caused by whether the infraspecies of TGBIF is within or equivalent to TEdl-Other; or that TEdl Theobroma cacao ssp. leiocarpum Bernoulli Cuatrec is within or equivalent to TGBIF-Other. At this point, it is difficult to discern which PW yields a more reliable merged solution. Expert opinions are needed to proceed further in this case.

Figure 5. Input alignment (top) and output Possible World (bottom) for TEdl and TITIS.

Figure 6. Possible Worlds for TEdl and TGBIF
5.0 Discussion and Conclusion

In this paper, we have presented a framework to reconcile taxonomies from different sources in BHL. Specifically, we have conducted six pairwise taxonomy alignments on four different taxonomies (ITIS, USDA Plants, EoL, and GBIF). As shown also in prior research (Franz et al., 2015, 2016; Cheng et al., 2017; Cheng and Ludäscher, 2019), a logic-based approach to taxonomy alignment can be used to align and merge different taxonomic perspectives into a solution that makes hidden relationships explicit. In this paper, the merged Possible Worlds of the alignments mainly serve as a subspecies grouping mechanisms that allow users to identify which taxonomies are more granular than the other, as illustrated in 4.0.2 and 4.0.3. The result for 4.0.1 also partially suggests that the PW may serve as a name disambiguation mechanism, where the grey boxes that groups equivalent terms shows that *Theobroma cacao* ssp. *cacao* L may be synonymous with *Theobroma cacao* subsp. *cacao* L. Further, while reserving a residual category “Other” to “designing for change” (Tennis, 2012) in classifications is usually considered a good practice, our results in 4.0.2, 4.0.3, 4.0.4 exemplified that the PWs will be partially influenced by the “residual category”. The children nodes may be classified into the “Other” category and creates ambiguities for the alignment results.

Moreover, our framework tries to minimize the information overload during the alignment process in these aggregated databases. Interoperability endeavors such as taxonomy alignments, cross-walking, or ontology mapping rely substantially on human decisions, especially when the topic of alignments is domain-specific. Experts of a domain asserts what kind of relations a concept in taxonomy A has with taxonomy B. In this paper, we attempted to reduce expert involvement at the beginning stage by using semi-automatic alignment process and incorporating existing external taxonomies. This is not to say that these taxonomies are the ground truth of how *Theobroma cacao*’s taxonomy should look like, nor that the species names aligned are the absolute answers for equivalencies (they are a lot of times not equivalent due to evolving semantic changes (Franz et al., 2016). Rather, these merged solutions are serving as interim knowledge organization systems pending to be further scrutinized by biodiversity and taxonomy experts in the future. Given the large amount of data in aggregated databases such as the Biodiversity Heritage Library, we believe using this approach to establish a minimal viable knowledge product first can be helpful for further efforts.

In future work, we plan to extend this study mainly by (1) employing this framework to generate result for more species; and (2) implementing the merged PWs as a new species information representation structure that can be used alongside BHL and assessing its retrieval effectiveness. In the opening scenario, Charlie’s searches actually further reveal that similar looking names in BHL such as *Theobroma cacao* L and *Theobroma cacao* Linnaeus yield identical search results with 2,733 records. This may suggest that BHL has an internal infrastructure that recognizes and organizes keywords together. We also hope to explore this and to incorporate parents (genus, family, order, or class-level), siblings (other species within the same genus), synonyms, and common names into the taxonomies to form more comprehensive species hierarchies in our framework. Ultimately, we hope to continue conversations with BHL on improving the practices of organizing species names and name reconciliation services.

Acknowledgement
This project is supported by the Center for Informatics Research in Science and Scholarship (CIRSS) at iSchool, University of Illinois at Urbana-Champaign. The first author
would also like to acknowledge the LEADS-4-NDP fellowship program, Dr. Jane Greenberg, and Steven Dilliplace for their ongoing support of this work.

References
Appendix 4.3. Result for $T_{\text{GBIF}}$ and $T_{\text{ITIS}}$ (Top: input taxonomies; Bottom: PW)

Result for $T_{\text{GBIF}} - T_{\text{USDA}}$ (Top: input taxonomies; Bottom: PW)
Abstract:
The concepts of representation and misrepresentation are recognized and used within Knowledge Organization. Although it is easier to find works related to the term representation in the literature, there are not many works that address the term misrepresentation from a conceptual perspective. Misrepresentation, as opposed to representation, has a negative connotation and it is considered a problem in Knowledge Organization. Despite the fact that misrepresentation could appear in any kind of knowledge organization systems, or that some systems have historically been criticized for the lack of representation of certain sectors of users, the term as such, has gained strength in recent years after the appearance of some theories that support more local systems instead of universal systems. In this work the concepts of representation and misrepresentation will be approached from a conceptual perspective and will be related to other terms within Knowledge Organization, such as bias. The relationship between positive bias or slant and representation, and negative bias as a form of misrepresentation. These relationships will be approached from the perspective of warrants, especially cultural warrant.

1.0 Introduction
The purpose of this paper is to provide an overview of the relationship between the terms representation and misrepresentation and their relationship with the term bias, specifically the terms positive bias and negative bias.

Although these terms are recognized and used within Knowledge Organization literature, it is not easy to find an approach from a conceptual perspective and precise terms’ definitions. Dubuc (1999, 33) claims that “the situation in which the terms are found condition the concept at the communication field to such an extent that [even] the same concept may receive different names or labels depending on the specialty in which it is used”. This also occurs the other way around, the same denomination may have more than one concept.

“[A] concept exists within personal knowledge structures, in one or more minds. The concept can be represented by a symbol, such as a word or string of words, that may be uttered or recorded (e.g. written). The symbol thus indirectly represents the referent. The same referent may give rise to varied concepts in different minds. The same concept may be represented by different symbols.” (Vickery 1986, 146)

As Gutiérrez Rodilla (2005, 10) states “Language is, […] a constitutive part of science. Therefore, it is impossible to learn science without knowing the language in which that science is expressed and without knowing how to correctly interpret its speech.” Therefore, it is necessary to use the exact terms in a specific field to assure a clear and coherent communication, avoiding ambiguities and misinterpretations.

The structure of this paper is as follows. In the following section the concepts of representation and misrepresentation will be developed from a conceptual approach. In the section titled ‘Word vs. term’ a terminological perspective of the terms will be provided. Pointing out the main variation between word and term will be beneficial to have a better understanding of the differences between the conceptual approaches.
Principally, the difference between the word ‘misrepresentation’ and the term ‘misrep-resentation’.

In the section after this, the terms representation and misrepresentation will be presented from the term bias point of view. In particular, according to negative bias and positive bias.

After that, the relationship among the above-mentioned terms will be discussed in relation to warrants and hospitality, more precisely cultural warrant and cultural hospitality.

2.0 The concept of representation and misrepresentation

The way in which the terms representation and misrepresentation appear in the professional literature about Knowledge Organization is very singular.

The term representation is commonly used in syntagmatic forms such as subject representation (Milani, Guimarães, and Olson 2014; Olson 2002), knowledge representation (Milani and Guimarães, 2011), among others.

In the case of subject representation, the meaning refers to topic representation in catalogues through indexing and classification. In other words, how the work’s content is represented through topics or descriptors from indexing and class numbers or notations from classification. According to Olson (2002, 3) they are “the key to subject access”.

In the book *The Power to Name* (Olson 2002) the author develops the representation concept through the action of assigning names to topics. She chooses the expression ‘naming’ because this action reflects a conscious will at the moment of representing concepts. She says “I choose the word naming because it connotes the power of controlling subject representation and, therefore, access.” (Olson 2002, 4)

By preferring a name (terms or words) to represent a concept, an identity to that concept is being established. This identity is biased, it has a way of recognizing and observing that reality.

Knowledge Representation could refer to either the name of an autonomous discipline from Knowledge Organization (Giunchiglia, Dutta, and Maltese. 2014) or a subordinate part of Knowledge Organization (Barité et al. 2015; Dahlberg 1993).

Giunchiglia et al. (2014, 47) say that through ontologies, Knowledge Representation “provides a more expressive representation and query language, able to codify and automatically query such knowledge”. On the other hand, in Barité et al. (2015, 136) the meaning of Knowledge Representation is described as:

“The group of processes of notational or conceptual symbolization of human knowledge in the field of any discipline. Knowledge Representation includes Classification, Indexing and the group of computer and linguistic aspects related to the symbolic translation of knowledge.”

All the cases mentioned above share the semantic background of the concept representation. Representation can be understood as “the description or portrayal of someone or something in a particular way.” (Representation 2019). Its meaning implies presenting again or presenting in a different way, for instance a concept.

In all the cases, the use of the term representation in Knowledge Organization has a positive connotation, i.e., a determined concept or phenomenon is correctly represented.

On the contrary, the term misrepresentation appears in an isolated way and not in a syntagmatic way. It usually accompanies the term representation in any of its forms
The use of the term misrepresentation is so linked to the use of representation that even its origin seems to be a consequence of the first. For a better understanding of this, it is necessary to explain the differences between word and term.

3.0 Word vs. term

Although at first word and term seem to be the same, they present some differences between them as far as their concepts and uses are concerned.

“A word is a unit described by a set of systemic linguistic characteristics and endowed with the property of referring to an element of reality.” (Cabré 1999, 25, in translation). While a term is “a unit of similar linguistic characteristics, used in a speciality domain. From this point of view, a word that is part of a specialized field would be a term.” (Cabré 1999, 25)

Thus, the principal difference between words and terms is their use field. Speakers are the ones who establish the label ‘word’ or ‘term’ based on different contexts and through their use. Pearson claims:

“While we accept that there are indeed differences between words and terms, we find that, without human intervention, it is not possible to use any of the proposed definitions of term as a means of distinguishing between terms and words. This is because terms very often look the same as words and frequently not only look the same as words but can also function as words, albeit in different circumstances.” (Pearson 1998, 8)

There are different definitions and ways of differentiating the concepts of words and terms. Cabré (1999), for example, mentions 4 situations in which words can be differentiated from terms:

a) by its users: words are used by any speaker of that language, while terms are used by the specialists or experts in a certain knowledge.

b) by the situation in which there are used: words are used in any form of communication, while terms are presented in more formal channels of communication.

c) by the topic they represent: terms usually refer to concepts within a specific field, while words are used to refer to wider variety of meanings.

d) by the kind of speech in which they usually appear: words are used in any type of speech, while terms usually appear in specialised speeches through their diffusion channels.

In this way, the expressions which belong to a specific field and which are used to refer to a particular concept within a particular domain are considered terms.

There are different ways to create terms and word. One of them is when a word or term moves to the other’s category. In other words, words turn into terms or terms become words. These processes are called terminologization (Gutiérrez Rodilla 2005) and de-terminologization (Meyer and Mackintosh 2000).

As Gutiérrez Rodilla states:

“the terms enjoy great mobility, both horizontally - that is to say, they move from one area of knowledge to another, with the same or different meaning - and vertically - even the most highly specialized can become words used daily by all speakers.” (Gutiérrez Rodilla 2005, 29)

It seems natural to consider this process as the way in which representation has been coined in the Knowledge Organization terminology. As it was explained in the section
before, the word and term ‘representation’ keep the same semantic background. However, this does not happen with misrepresentation.

The definition of the word misrepresentation is the following: “the act of deliberately giving false information to someone, especially in order to persuade them to enter into a contract, or a statement giving false information.” (Misrepresentation 2019)

Despite the fact that both term and word have negative connotations, the definition of the word ‘misrepresentation’ implies a willing and conscious action. Misrepresentation is being created to persuade somebody or pursue something deliberately.

Nonetheless, this not always the case in Knowledge Organization. A knowledge organization system could be misrepresented even if the author did not have the intention of doing it. The reason of this could be the lack of knowledge of other realities or ways of presenting concepts, the way in which this author perceives reality or the predominant way for this author, and finally just because it is representative for a community but not for another.

Another hint could be that, as it was mentioned before, the term misrepresentation usually appears together with the term representation. If the prefix ‘mis-’ in English is analysed, it has 3 different senses: “1 bad or badly; 2 wrong or wrongly; 3 used to refer to an opposite or the lack of something” (Mis- 2019).

The term misrepresentation covers the 3 aspects in Knowledge Organization. A knowledge organization system can present misrepresentation through bad representation, a wrong representation or a lack of representation.

Taking all this into consideration, misrepresentation as a term does not seem to come from a terminologization process as it happens with the term representation. It could be inferred that misrepresentation is created in contrast to the term representation.

4.0 Bias as a form of representation and misrepresentation

There are several authors who recognize the existence of bias in knowledge organization systems (Higgins 2012; Mai 2010). Despite this, there are few of them who identify the different aspects of the term and connect it with the representation and misrepresentation concepts. As Broughton says:

“bias is said to exist when a controlled vocabulary contains an unduly large number of terms reflecting the ideas, interests or positions of a particular sector or field, or when terms relevant to another sector or field fail to appear. This may occur because the language of a particular group is preferred.”

(Broughton 2012, 256)

From the concept of bias two situations can be established. One of these situations indicates when a way of representing reality is relevant and beneficial for a certain community. This bias is called positive bias (Colombo 2015; Colombo and Barité 2015) or slant (Guimarães 2017).

The other situation happens when a bias does not represent the ideological, cultural peculiarities and fails to represent the concepts, in some cases reaching prejudice. This aspect is called negative bias (Colombo 2015; Colombo and Barité 2015).

Using Virtual Reality domain as an example of the use the concept of bias in relation to representation, Brey (1999, 12) says “when a VR application favours certain values or interests over others due to its choices in representation, it may be said that the model makes use of biased representations.”
Consequently, a knowledge organization system with a certain representation or representing a specific group or way of thinking, is a biased system from that point of view or form. On the other hand, a system that has misrepresentation as a result of a void in the representation or due to the fact that the concepts are not represented in a proper way, it has a negative bias.

Following the example of Virtual Reality:

“When a VR application fails to uphold accepted standards of accuracy by representing features as real that by such standards cannot justifiably be held to be present in reality or by failing to represent features that ought to be present in the application, we may say that the application misrepresents reality” (p. 11) (Brey 1999)

Whereas Brey identifies both cases, ‘bias representation’ and ‘misrepresentation’ as “two types of representational failures or shortcomings” (Brey 1999, 12), this does not happen in Knowledge Organization.

In the last years, Knowledge Organization has focused more on local systems instead of pursuing universality. Hence and according to this, it may be helpful to have knowledge organization systems positively biased. In Mai words “while modern classification aims at representing the universe of knowledge, postmodern classification aims at providing a pragmatic tool for specific domains.” (Mai 2004, 39)

The issue is to decide for which community sector the bias is representative and for which it is not and how to detect these particular characteristics. “Verifiable misrepresentation requires that there are unambiguous, shared standards of accuracy in place according to which judgments of misrepresentation can be made.” (Brey 1999, 11)

One of the ways to determine the cultural aspects of a sector is through the cultural warrant.

5.0 Cultural warrant and cultural hospitality in relation to representation

A form of representation of a certain group in a knowledge organization system can be determined from the warrants. Depending on the approach of what needs to be represented or the bias that is sought to be obtained, it could be literary warrant, academic warrant, cultural warrant, among others.

Literary warrant is based on the documents while academic warrant is based on the opinion of experts and cultural warrant, in particular,

“means that any kind of knowledge representation and/or organization system can be maximally appropriate and useful for the individuals in some culture only if it is based on the assumption, values, and predisposition of that same culture. Conversely, if a system is not based on those assumptions, it will be appropriate and useful to some lesser extent for the individuals in the culture” (Beghtol 2002, 511)

It is important to bear in mind that these warrants can be combined and are not exclusive to each other. To read more about warrants see (Barité 2018).

In relation to the concept of ‘cultural warrant’ is the concept of ‘cultural hospitality’. The term ‘cultural hospitality’ is a deviation of the term ‘hospitality’.

The concept of hospitality implies that a knowledge organization system is capable of introducing a new concept or term into its structure. The system must provide tools not only for the inclusion of an element, but also for establishing relationships between them, generating more permeable and not so rigid systems. Cultural hospitality in particular “means that a knowledge representation and organization system can ideally accommodate the various warrants of different cultures and reflect appropriately the assumption of any individual, group, or community.” (Beghtol 2005, 905)
In any case, the system has to provide a clear mention about not only how to introduce new concepts, but also for whom the system is set, or in other words, for which user community is more representative.

6.0 Conclusion

All things considered, it can be observed that the terms ‘representation’ and ‘mis-representation’ have a close relationship with the terms ‘positive bias’ and ‘negative bias’. In this context, bias is considered as a form of representation. What is more, it is not possible to think about representation without thinking about cultural warrant as a means to ensure a correct and better representation for each situation.

References

Barité, Mario, Stephanie Colombo, Amanda Duarte Blanco, Lucía Simón, Gabriela Cabrera Castromán, Maria Luisa Odella, and Mario Vergara. 2015. Diccionario de Organización del Conocimiento: Clasificación, Indización, Terminología (6th ed.). Montevideo: CSIC.


Giulia Crippa – University of Bologna (Ravenna Campus), Italy
Andre Vieira de Freitas Araujo – Federal University of Rio de Janeiro (Rio de Janeiro), Brazil

Order of Knowledge, Selection and Bibliographical Tension in the 16th Century
Between Gesnerian Universality and Possevinian Anti-Heretism

Abstract:
Our work is guided by the construction of the genealogy of knowledge organization (KO). The plan of our discussion is that which is linked to the constitutive elements of modern bibliographical principles. In the history of cultural circulation, after the invention of the printed book, at least two interpretative models of the function of the culture and the representation of knowledge in society can be observed. On the one hand, the lay principle is developed according to which man finds his dignity in the rational search. The antagonistic side, on the other hand, develops the dogmatic view of those who consider themselves to be representatives of unshakable certainties that, from the point of view of a superior good, impede the freedom of choice of individuals. Among the first great representatives of these two models, we can see how Conrad Gesner, in his secular-minded compilation, incorporates a first proposal, while Antonio Possevino, on the other hand, proposes a bibliography rigidly configured to support the Catholic Counterreformation spirit (Serrai and Sabba 2005). From a historical-bibliographic approach intertwined with the ongoing debates in KO, the objective of this study is to promote a brief comparative formulation between the gesnerian Bibliotheca and the possevinian Bibliotheca, from the selection and the bibliographical tension under the aegis of universality and anti-heretism that delineate the order of knowledge in the 16th century. The study indicates that a proposal of theological and moral control in the access to knowledge, made through bibliographical control, finds in Possevino one of its greatest representatives (Santoro and Orlandi 2006). In this context, knowledge is evaluated through the rejection filter of authors who do not adhere to Catholic dogmas, which is why the Index Librorum Prohibitorum is instituted, condemning authors such as Giordano Bruno, Copernicus, Galileo and own Gesner. On the other side of the reflection on knowledge and information between the 16th and 17th centuries is the view of authors whose roots lie in the more properly humanistic culture. Genealogically, in bibliographical terms, it is Gesner and his work that heads this view, which tends to be delineated as lay knowledge. In comparative terms, as stated by Serrai (1993), while Conrad Gesner builds a Bibliotheca, which was Universalis, with Science, Nature and Theology as its main roads, Antonio Possevino draws his own Bibliotheca, which was Selecta, as a map of knowledge protected, guaranteed, without danger, by orthodoxy and morals. In Gesner there is absolute certainty in science and in Possevino reappearance a recurring anthropological skepticism and suspicion about the value and innocence of science. Indeed, the selection and the bibliographical tension between the two Bibliothecae, under the aegis of universality and anti-heretism, become the key not only to understanding but also to the delineation of the order of knowledge in the 16th century. In the long-term perspective, they are horizons that reveal the historical relations between knowledge, its control, its access and its organization.

1.0 Introduction
Our work is guided by the construction of a genealogy of knowledge organization (KO), an archeology in which in each stratum, each epoch is revealed.

If a complex bibliographical repertoire in medieval libraries does not seem to be relevant, because relatively few manuscript materials are available, one has to go to Trefler and, especially, Gesner, to find a substantial difference in bibliographical principles, which can be accomplished through a genealogical study of bibliographical treatises, a study that allows hermeneutic potentialities, provided that the criterion for evaluating the relationship between what should be faced and the way to deal with it is followed,
that is: the relationship between problems in search of solution and solutions offered, making it necessary to accept the phenomenological polymorphism of library reality in its history, because any linearity and imposed coherence turns out to be false.

In order not to slip into conceptual misconceptions, it is necessary that the requirements of scientific explanation be limited to the observation of relationships and nothing else, relying on that single theoretical core represented by the individualization and functionalities of index and catalog relations that constitute critical matter and interpretative conditions of the Bibliography.

Bibliography represents information, making it necessary to recognize the existence of logics and mediation procedures via indexes and catalogs. In this sense, the plan of our discussion is that which is linked to the constitutive elements of modern bibliographical principles.

After the printed production of information in books, Bibliography acquires an essential role in order to be able to reformulate the library structure itself. In fact, the modern library reformulates itself because there is a change in the structures of knowledge and, therefore, in the logic of information organization, mainly through bibliographical and catalog production. The bibliographical structure, destined to become physical structure in libraries, is based, in our view, on a new settlement of the “parties” involved in the discussion about knowledge.

We talk of “parties” because the univocal voice of the medieval Christian world is already fragmented from the perspective of Renaissance humanism. The Renaissance period brought, as a result, the formation of Protestant churches in the religious field, as well as the empirical-experimental foundation of the first claims of modern science.

The result of all this is the strong reaction of the Catholic Church, which is reformulated at the Council of Trent (ended in 1563), from which the characteristics of the actors of a new dialectic of knowledge emerges. As Balsamo (2017) writes, there is, in fact, a genealogy on the basis of statements about organization and access to information, which are based on the same principles we ask ourselves about today: for what and for whom does one select, order and allow access to knowledge?

In the history of cultural circulation, after the invention of the printed book, at least two interpretative models of the function of culture in society can be observed. On the one hand, the lay principle is developed according to which man finds his dignity in the responsible and rational search for truth, in the attentive and inexhaustible search for an understanding of the reality around us and of which we are constitutive elements. On the opposite, the antagonistic side develops the dogmatic view of those who consider themselves to be representatives of unwavering certainties that, from the perspective of a superior good, impede individuals' freedom of choice, as orthodox interpretation is guaranteed by official institutions, whose other task is to control its dissemination.

Among the first great representatives of these two models, we can see how Conrad Gesner, in his “universal” compilation of secular spirit (Bibliotheca Universalis), embodies the first model while, on the opposite side, Antonio Possevino proposes a bibliography (Bibliotheca Selecta) rigidly configured to support the Catholic Counterreformation spirit (Serrai and Sabba 2005).

From a contemporary perspective, we can build an imaginative exercise and situate the aforementioned contrast as a kind of precursor to the tension between “universal”
and “domain-oriented” KOSs, which have been objects of relevant discussions in the last decade, notably between Hjørland and Szostak.

For Hjørland (2017), for example, the KOSs would be based on social relations established by specific domains, which is in line with the historical experiences that support our study.

Naturally, in the context of this work, we are positioned in the 16th-century timespace in which informational, philosophical, social and cultural particularities must be rigorously considered and relativized when observing how knowledge was produced in specific “domains” and specific “discursive communities”.

From the historical-bibliographic approach intertwined with the ongoing debates in KO, the objective of this study is to promote a brief comparative formulation between the gesnerian *Bibliotheca* and the possevinian *Bibliotheca*, from the selection and the bibliographical tension under the aegis of universality and anti-heretism that delineate the order of knowledge in the 16th century.

2.0 Conrad Gesner and *Bibliotheca Universalis*

Conrad Gesner (1516-1565), as is already known, was a Swiss scholar, scientist and bibliographer. His education took place in many different cities like Zurich, Bourges, Paris, Montpelier, Basel and Strasbourg.

Gesner had Ulrich Zwingli (1484-1531) as spiritual guide, intellectual reference, tutor and economic support. Zwingli had brought together and harmonized the principles of Christian Theology, with the exercise of human reason and the intellectual heritage of classical pagan civilization. In this perspective, in the center of Zwingli’s thought was God conceived as truth and supreme good, who had distributed to all the possibility of accessing truth and salvation from the moment of creation and not, as was believed in both Catholic and Protestant, thanks to the subsequent incarnation (Sabba 2012).

The "Zwinglian theological vision" was adopted by Gesner, a particular current of the Protestant Reformation - distinct from Lutheranism and Calvinism - which had been marked and guided precisely by head of the Zurich Church (Sabba 2012).

A quintessential Renaissance “polymath”, Gesner had the ability to articulate and discuss numerous areas of knowledge, publishing books on multiple topics such as linguistics, medicine, theology, botany, zoology, paleontology, mineralogy and bibliography.

Bay (1916, 54-55) summarizes Gesner's relationship with his time and knowledge: “Gesner belonged to a period in the history of science distinguished for magnificent scholarship and elaborate method. His period of development and maturity was the ripening period of the Reformation. It was no rare occurrence that a man made himself master of the essentials of all knowledge thus far accumulated. [...] Gesner had that peculiar ingenium which marshals both wisdom and knowledge”.

Conrad Gesner's intellectual maturity and methodological rigor is consolidated in his work *Bibliotheca Universalis*. A seminal work in the field of Bibliography, *Bibliotheca* consists of an alphabetical-nominal part called *Bibliotheca Universalis* (1545) (Figure 1) and *Pandectae* (1548, 1549), which is systematically ordered on the semantic content of the works.

*Bibliotheca Universalis* (1545) is an alphabetical-nominal catalog that lists 5031 authors from around 15,000 works in Latin, Greek and Hebrew. It is organized in alpha-
Bibliotheca's authorial scope includes erudite and non-erudite authors. Therefore, for Gesner, everyone should be remembered; therein lies a relevant aspect of his project, which is the possibility of giving voice to unknown authors, which makes Bibliotheca a device of wide dissemination.

It is interesting to note Gesner’s refusal to discriminate, upon finding his humanist stance, leaving the reader to evaluate and even judge the sources. As Gesner says: "I wanted to report so much, but I left the selection and judgment of the books to others" (Gesner 1545, 3v).

Gesner’s deliberate choice to treat all authors as worthy of memory points to a criterion adopted by Gesner: a supposed documentary impartiality.

According to Serrai (1990, 82):

“With this criterion of absolute documentary impartiality, Gesner planted another pillar that supports the techniques and ethics of bibliographic disciplines: the registration, organization and preservation of documentary memories cannot be subordinated to any ideological preference".
For Gesner, the bibliographical operation should not be subject to restrictions or censorship, but, considering that Bibliotheca could also be used by inexperienced people, Gesner gives advice, guidance and warnings in relation to poor quality works.

From the point of view of the organization of knowledge, it is worth remembering that Gesner proposes, in the second part of the Bibliotheca, called the Pandectae (1548), a classification system that expands the seven liberal arts of Medieval tradition to the categories of complementary subjects of interest to the Renaissance scholars, constituted by 21 classes or partitions.

Gesner elaborates the Pandectae with the following classification structure for books:

In his point of view of nature and his choices related to the taxonomy, Gesner reveals a scheme that seeks to contemplate the totality of orders: natural and artificial; of things and of sciences.

### 3.0 Antonio Possevino and Bibliotheca Selecta

Antonio Possevino was born in Mantua, a small city in the north of Italy, in 1533. Mantua was, at that time, an important court, governed by the Gonzaga family. Possevino went to Rome in 1550 to study and, in 1554, became Cardinal Ercole Gonzaga’s secretary, working, at the same time, as teacher for the future cardinals Francesco and Scipione Gonzaga. In 1559 Possevino entered the Society of Jesus, a turning point in his life. From then on, he became a dedicated preacher against heresy and spent many efforts to try to solve theological and political matters with northern and eastern European countries. He travelled to France, Sweden, Poland, Russia, Hungary, Romania and Moravia, looking for reconciliation where there were schisms proposed. He died in Ferrara in 1611 (Serrai 1993).

The textual structure of Possevino’s Bibliotheca (Figure 2) is based on the treatise modality, accompanied (as we said) by tables and authors. Books are divided into Holy Scriptures, Positive Theology, Scholastic Theology, Catechetical Theology, Practical Theology, Clergy, Heresy, Philosophy, Law, Medicine, Mathematic, Music, Architecture, Cosmography, Geography, History, Poetry, Oratory and Miscellaneous (Serrai 1977).

The title of the work Bibliotheca selecta qua agitur de ratione studiorum in historia, in disciplinis, in salute omnium procuranda explains the relationship with the Ratio Studiorum, a pedagogical system established by the Jesuits for their educational centers, which will be published in 1599. Faced with the universality and impartiality of the information offered by Gesner’s Bibliotheca, a list of authors sorted alphabetically and by topic in the Pandectae, Possevino’s objective is to propagate Christian doctrine to remove heresies and annihilate the schism. To this end, the curriculum proposal aims to provide for an each individual, based on their conditions and social status, the indications of the authors and the appropriate readings, by the children of the princess, oriented
towards civilians, ecclesiastics, diplomacy, passing through the nobility, even the lowest classes.

Alongside the indications for the *dispositio* and the good conservation of the volumes and the catalogs of the works and the nomenclatures of authors, the *Bibliotheca Selecta* offered indications for the *emendatio* and the *expurgatio* of those works that would otherwise have been prohibited, in addition to committing to refuting works and authors already listed. So Possevino, in the "model library", felt the need to return to opposition to certain publications, positioning his *Bibliotheca* not only as a mirror but also as a complement to the *Index Librorum Prohibitorum* (Balsamo 2017).

On the basis of Possevino's work, the Italian monastic libraries were purged at the end of the sixteenth century while, with regard to Rome, the libraries of the great cardinals were recipients of Possevino's censorship program (Serrai 1993). It is also true that in the *Bibliotheca Selecta* the primary interlocutors were princes and nobles who possessed rich libraries. All the rest of the work turned to the Jesuitic Order and the same *Bibliotheca* was initially conceived by Possevino as a bibliographical and at the same time pedagogical work to be destined primarily for the principles, considered on the one hand as the users of the Jesuit institution and, on the one hand other as the defenders of
Christianity. In the libraries of princes, therefore, not only printed books but also manuscript codes had to be subjected to rigorous censorship.

A prescriptive bibliographical canon, such as the Bibliotheca, necessarily has a closed and imposed character, based on the Counterreformation Catholic doctrine. It is articulated following a hierarchical scheme that begins with the Divine History, then Positive Theology. The Scholastica Theology follows: that is, the interpretation of the sacred writings according to the teaching of the Church. Next comes the Theology practice as a spiritual direction of consciences, and Catechetic Theology, oriented to pedagogical activity, with the establishment of a whole curriculum for the school. Eleven of the eighteen books are dedicated to all this part. The autonomy of human science is questioned, because all sciences are included in Divine History.

The Bibliotheca classification scheme is radically opposed to Gesner's Pandectae, which began with the Trivium and Quadrivium, to end with Theology in the twenty-first book. The Bibliotheca represented a guide to safe, guaranteed knowledge, without danger for Catholic orthodoxy. The work is organized in two volumes. The first, dedicated to Pope Clement VIII, is divided into eleven books: the first five lay the foundations of Christian education on the Scriptures and on Theology, while books VI-XI provide the cultural tools for evangelization of the world by reformed Christians to the inhabitants of the Indies, in view of a Catholic "conquest" or "reconquest". The second volume, dedicated to Sigismund III, King of Poland and Grand Duke of Lithuania, consists of six books (XII-XVIII) in which the different disciplines (law, philosophy, medicine, mathematics, architecture, geography, history, poetry, painting and rhetoric) are presented in descending hierarchical order and dependent on Theology, in controversy with their alleged autonomy.

Possevino's Bibliotheca can be used to draw a balance of early modern Catholic culture. Quotes and oversights, genuinely known texts and second-hand quoted texts, corrections and errors, convictions and censures, autobiographical references and (not always explicit) positions on current problems are a fertile ground that still awaits investigation.

It’s interesting to notice that his suggestions for organizing a physical library differ from the bibliographical scheme, as Serrai indicates (1977, 79).

4.0 Order of knowledge, selection and bibliographical tension between Bibliotheca Universalis and Bibliotheca Selecta

On the other side of the reflection on knowledge and information between the 16th and 17th centuries is the view of authors whose roots lie in the more properly humanistic culture. Genealogically, in bibliographical terms, it is Gesner and his work that heads this view, which tends to be delineated as “lay” knowledge. The idea of "universality" present in the Gesnerian work does not point to a generic totality, but to the possibility of access and appropriation of books and manuscripts by the learned community. If, at the time, this community is made up of scholars, at no time are there obstacles to its expansion.

The ideological question involved, though, should be linked not to the simplistic opposition of religion versus science, or past versus future. The two models of interpreting the world are far more complex than this. As Serrai (1993, 717) points out, while Gesner's bibliography was Universalis, Possevino’s is Selecta, which means that the
sources of the first are broad and that the sources of the second are strictly chosen. While Gesner looks for the widest compilation of works covering all the fields of knowledge (a compilation limited only by the three chosen languages), Possevino creates a bibliography aimed at education, study, reference inside the domain of orthodox theology. Gesner, as Balsamo states, considers bibliography “an essential tool for achieving knowledge and ‘communicating’ it to others”, being “an invitation to share in further research” (2017, 30). The core of Possevino’s Bibliotheca consists of “mapping” all the knowledge fields through texts, tables and the list of authors that wrote about each specific field, including those not acceptable by the reformed catholic doctrine, being not recommended. Gesner builds a catalogue meant to be not only the state of the art of knowledge, but also an effort to offer traces of all the previous culture. Possevino, on the other side, serves the Church purpose to reestablish its primacy as knowledge “broker”, interpreter of the right doctrine that ascend to the Divine. This way, the main purpose of Bibliotheca Selecta was to be “prescriptive bibliographic canon which would serve as a tool for imposing ideologically correct works on all who engaged in studies or research” (Balsamo 2017, 46-47).

This opposition, stated by the very title of Possevino, should be smoothed by the relevance of Gesner as a source for his bibliography, although Gesner is cited by the author both in private letters (see Serrai 1993, 113) and in the Bibliotheca Selecta (also cited by Serrai, 1993, 717 and 720, referring the first to the Preface to Bibliotheca Selecta and the last one to the Apparatus Sacer).

One should consider that after he entered the Jesuit order, in 1559, Possevino had to contend with the relevance of the Bibliotheca Universalis throughout the European intellectual circles, even though included already in the first edition of The Index Librorum Prohibitorum, in 1564.

The Catholic Church needed a modern structure for sustaining its authority, undermined by such a complete catalog as the Bibliotheca Universalis, whose contents trespassed the boundaries established by the Counterreformation. Gesner’s Bibliotheca had become so relevant that the Church had to choose whether to accept it or to lose the competition to sustain its knowledge authority.

Possevino, in his Bibliotheca, forged a strategic tool based on the impossibility of universality: this would imply the acceptance of Protestant authors, while his interest focused only within the domain of Catholic knowledge.

In order to better explain this dialectic, we offer an example taken from the disciplinary position of Catholicism in relation to artistic production. Clearly, the post-conciliar Church establishes rules for the realization of religious images, for which the didactic function stands out. That art conceived in this way had a brief life, is evident in its rapid evolution to the emotional appeal of Baroque representations that, even so, maintain their theological rigor, expressed by an effective rhetoric. We have already talked about Possevino's role in the elaboration of Counterreformation bibliographical catalogs, and it is worth mentioning that he dedicated himself to the bibliography related to art, in his Tractatio de Poesia et ethnica, humano et fabulosa collata cum vera, honesta et sacra, from 1595 (Possevino, 1971). Just like his Bibliotheca Selecta, Poesia et Pintura offers the rigidly delimited model of the Counterreformation doctrine, a model that becomes an instrument of close control applied to bibliographic information and the circulation
of books, aimed at the construction, “on the documentary level, of a collective memory selected according to a specific pedagogical program” (Balsamo 2017, 55).

What we want to highlight here is that Possevino, a religious scholar and bibliographer, selects a set of authors and books not dedicated to techniques, but rather to morality in painting and sculpture. Possevino, is not an artist, so he expresses moral concerns on art, selecting those authors that “deal with this issue from a theoretical point of view, and not a practical one, as other art bibliographers were doing at the time, in order to structure the meaning of the object of art. This way, he offers titles that move away from the technical domain” (Crippa 2018, 76).

Returning to the scheme of the actors of the dialectic of knowledge of the time, one can observe, on one side, the proposal of a theological and moral control in the access to knowledge, which is accomplished through bibliographic control, which finds in Possevino one of its greatest representatives (Santoro and Orlandi 2006).

On the other side a libertarian, bourgeois matrix thinking develops, proposing a “universal”, secular access to knowledge. If we rely on this dialectic between the two models, it is appropriate here to offer a proposal for the individualization of their characteristics, focusing on the library as a public service that provides all the tools for study and information. We thus identify the current of thought linked to the post-conciliar vision, in which the control by the ecclesiastical institution of knowledge through its rigidly controlled administration and dissemination is placed as its basic principle. Perhaps, it should be remembered, once again, the role played by the new religious order of the Society of Jesus, an order specifically created to support the decisions of the Counterreformation. In any case, knowledge is thought the filter of rejection of authors who do not adhere to Catholic dogmas, which is why the Index Librorum Prohibitorum is instituted, condemning authors such as Giordano Bruno, Copernicus, Galileo and, not surprisingly, Gesner.

5.0 Considerations

In comparative terms, as stated by Serrai (1993), while Conrad Gesner builds a Bibliotheca, which was Universalis, with Science, Nature and Theology as its main roads, Antonio Possevino draws his own Bibliotheca, which was Selecta, as a map of knowledge protected, guaranteed, without danger, by orthodoxy and morals.

In Gesner there is absolute certainty in science and in Possevino reappearance a recurring anthropological skepticism and suspicion about the value and innocence of science.

Gesner turns to scholars and elaborates for them the indices of the cultural heritage of all humanity, creating a mediating instrument for documents and monuments, without sectarianism and bias. Possevino, on the other hand, works in reverse, under the threat of Protestant advancement, and cannot be faithful to the principle of universality: he prepares a guide for those who, as Catholics, must be safeguarded, tutored and protected (Serrai and Sabba 2005).

Gesner and Possevino became emblems of two cultural worlds and gave favorable conditions to the development of science and civilization.

Gesner promoted a bibliographical selection based on criteria of intellectual, scientific and philological rigor and developed a rigorous method of bibliographical nature. But a substantial difference with Possevino lies above all in the greater
conceptual breadth of Zwingli's ideological system, in which the Gesnerian culture was implanted, compared to that of the narrow doctrinal armor that marked the ideology of the Catholic Counterreformation (Serrai and Sabba 2005).

In Modern Europe, Gesner and Possevino promoted bibliographical ruptures, of a Protestant and Catholic nature, respectively, that affected significantly the forms of production, organization and mediation of knowledge.

Indeed, the selection and the bibliographical tension between the two Bibliothecae, under the aegis of universality and anti-heritism, become the keys not only to understanding but also to the delineation of the order of knowledge in the 16th century. In the long-term perspective, they are horizons that reveal the historical relations between knowledge, its control, its access and its organization.

References
https://doi.org/10.3931/e-rara-16206.
Possevino, Antonio. 1593. Societatis Iesu Bibliotheca Selecta Qua Agitur de Ratione Studiorum in Historia, in Disciplinis, in Salute Omnium Procuranda. Romae: ex Typographia Apostolica Vaticana. https://books.google.dk/books/about/Antonii_Posseuini_Societatis_Iesu_Bib-
lio.html?id=H8h6SUORc8IC&redir_esc=y
Crafting a System for Knowledge Discovery and Organisation
A Case-Study on KOS for a Non-Standard German Legacy Dataset

Abstract:
This paper describes a case-study developing a knowledge organisation system (facet thesaurus) on the example of a non-standard German language legacy dataset, DBÖ [Datenbank der bairischen Mundarten in Österreich / Database of Bavarian Dialects in Austria]. A particular focus is placed on the 109 original data collection questionnaires contained in the collection, which are understood as an entry point to the entire collection. Here they serve as a case-study to demonstrate the process, which may be extended to the remainder of the collection. Ranganathan (1933, 1967) created the first faceted scheme - Colon Classification - to classify books in libraries. Faceted classification has also been used to assist automated search and retrieval of information (Prieto-Diaz 1991). According to Mills (2004), facet analysis has a very vital role for information retrieval and in the design of classificatory structures by the application of logical division to all forms of the content of records, subject and imaginative. The natural product of such division is a faceted classification. Building on these previous endeavours, we here introduce a facet thesaurus for eliciting and promoting access and navigability for the items in this collection, in order to make cross cutting topics accessible.

1.0 The aim and scope of the study & introduction
The aim of this study is to introduce a first approach towards creating a knowledge organization system (facet thesaurus) for a non-standard German language legacy resource (DBÖ)\(^1\), enabling transversal knowledge discovery. Here we present the rationale to our approach, the applied methodologies, and a concrete example on technology related terms in the form of a case-study that, in a next step, can be readily applied to other thematic areas within and beyond the collection.

Our undertaking is realised within the project exploreAT!\(^2\) and the wider framework of exploration space\(^3\) at the Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-OeAW). The exploration space is a digital and physical space offering opportunities for experimentation and innovation in the networked Humanities, and has since its establishment in 2017 also been listed as a best practice example for Open Innovation in the Humanities\(^4\). The project exploreAT! is a multidisciplinary endeavour with international collaboration partners (semantic technologies: Adapt Centre, Dublin City University, IE; visualisation tools: VisUSAL, Universidad de Salamanca, ES), with the general aim of opening the DBÖ collection for thematic exploration and exploitation.

---


\(^2\) https://www.oeaw.ac.at/acdh/projects/exploreat/

\(^3\) https://www.oeaw.ac.at/acdh/about/core-units/core-unit-4/

\(^4\) http://openinnovation.gv.at/portfolio/oeaw-exploration-space/
by means of semantic technologies and visual prototyping, as well as linking to other resources.

The DBÖ collection is large and rich (~3.5 million entries) and is composed of digitised data collection questionnaires, answers as well as excerpts of vernacular dictionaries and folklore literature. The data, having undergone several stages of digitisation, is to-date available in TEI/XML format and partly as a MySQL database. The questionnaire data (109 thematic questionnaires comprising around 17,000 individual questions) thus constitutes only a fraction of the DBÖ. The questionnaires originally pertained to a dictionary project aimed at capturing the German language spoken by local population from the early 20th century onwards in the area of the former Austro-Hungarian empire. Therefore, apart from being a rich linguistic, non-standard resource, the collection also captures a wealth of historic cultural information of everyday life, e.g. customs, religious festivities, food, traditional medicine, professions, songs, among others. The answers to the questionnaires’ questions thus follow a lexicographic structuring and are composed of headwords (lemmas), senses/meanings, sources, geographic location information (GIS), person information (authors, collectors, data typists), etc. In the context of the exploreAT! project, opening up the collection and questionnaires to new ways of exploration was initiated via lexical concepts, enabling the linking to other resources, such as Linked Open Data (LOD) (cf. Abgaz et al. 2018; Dorn et al. 2019). With this came the necessity for transversal knowledge searching across questionnaires and access to knowledge otherwise inaccessible or hidden. The main topics of the questionnaires reflect more or less detailed aspects of everyday life, which is also reflected in the quantity of questionnaires dedicated to a certain topic. Topics such as “movement” (Bewegung) (11 questionnaires), “wedding” (Hochzeit) (5 questionnaires), “tailoring” (Schneiderei) (4 questionnaires) or “baking bread” (Brotbacken) (3 questionnaires) have been queried extensively, while other topics (e.g. body parts, time, animals, school and education, plants, brewing) were covered with fewer questionnaires. Based on this content and the overall topics represented, we have chosen use-cases that would allow us to explore the data from perspectives relevant for cultural exploration as well as from the Digital Humanities perspective in general. Exploring the basic subject of “technology” across the collection, on the one hand, would allow us to provide a “historical” view of technology related terms and concepts from the time when the questionnaires were conceptualised (1912)\(^5\), while, at the same time, providing a noticeable contrast to what we understand by technology today. The other selected use-case deals with the topic of “food”, which is covered by specific questionnaires, but also transversally with specific, food related questions occurring across several questionnaires. In addition, food is one of the key aspects of mankind’s culture and thus particularly relevant not only from a historical perspective but also nowadays.

Navigating through such a vast collection is hard when dealing with purely term based search interfaces, and that is why we have chosen the faceted approach to organize the main concepts of the collection in a thesaurus, that would serve as the entry point for navigating through the questions, each one linked to one or (many) more answers. Relevant terms were chosen using its raw frequency in the questions. However, word

---

\(^5\) https://vawadioe.acdh.oeaw.ac.at/projekte/wboe/geschichte-des-wboe/
embedding models (Mikolov 2013a) were also used to explore semantic vicinities and to find similar concepts among the whole collection. Ultimately, we have achieved our goal to build a navigable interface providing access to the data fields in the collection of cultural heritage.

2.0 Theoretical background

In the scope of this project, two main knowledge representation techniques were used: faceted analysis and word embeddings. A brief overview of these is provided in this section below.

Faceted classification is a concept and a technique introduced by Ranganathan (1933, 1967) and later developed by the Classification Research Group (CRG) (Vickery 1960). A faceted scheme has several facets and each facet may have several terms, or possible values. A faceted classification scheme for wine, using (Broughton 2006) example, might include the facets (and terms) “grape varietal” (riesling, cabernet sauvignon, etc.), “region” (Napa Valley, Rhine, Bordeaux, etc.), and “year” (2001, 2002, etc.). According to Ranganathan, the process of choosing the facets is analytico-synthetic, that is, we first analyze the subject domain and then shape their compounding facets to adequately describe its characteristics and provide room for the concepts. That makes it a very powerful resource to describe and organize information. The facets need not be ordered, nor be of the same type, although they should be clearly defined and mutually exclusive (Broughton 2006). It was first devised for the classification of books in libraries (Ranganathan 1933), but was it subsequently adopted in information retrieval systems and search interfaces on the web (Prieto-Diaz 1991; Broughton and Lane 2000; Tudhope et al. 2006). Facet analysis has been used in the construction of information retrieval (IR) thesauri since the publication of the Information Retrieval Thesaurus of Education Terms in 1968 (Spiteri 2000; Barhydt, Schmidt, and Chang 1968). Spiteri (2000) states that there is no standardized model for the application of facet analysis to information retrieval thesauri and that national and international guidelines for thesaurus construction make minimal mention of the use of facet analysis. She presents, in his study, some critical arguments on how to evaluate the choices of facets for thesauri, and if these choices are coherent to the principles stated by both Ranganathan and the CRG. Ranganathan (1933) has proposed, influenced by the Brahman philosophy (Mazzocchi, 2013), the adoption of basic subjects and 5 characteristics of division used to derive facets: Personality (that we can interpret as “the things”), Matter (its characteristics), Energy (the processes), Space, and Time. His facets system has since the epithet "PMEST". The CRG, alternatively, preferred an ad hoc approach, and proposed that each subject area should be divided into categories that are appropriate to its nature. This latter approach was more suitable for our collection.

Word embeddings are one of the many powerful NLP techniques that have been developed in the past few years. To build semantic models out of large textual collections, we need to represent the semantic units into mathematical vectors. There are mainly two ways to construct this representation: using the simple bag of words model (Zhang et al. 2010), where each word is represented by a specific vector in a huge multidimensional space; and using word embeddings models, such as Word2vec (Mikolov et al. 2013a; Mikolov et al. 2013b), where each word is represented by a linear combination of a smaller set of dimensions or vectors. These “basic” vectors for word
composition are obtained using specific neural network architectures (e.g. “continuous bag-of-words” or “skip-gram”) (Mikolov, Yih, and Zweig 2013). These distributed vector representations of texts impressively capture syntactic and semantic aspects of concepts and their relationships. To generate these contextualized word representation models, it is necessary to feed the underlying neural networks with large corpora of texts of a specific language, so that the transitive relations between the concepts that co-occur in "windows" of contiguous words in a sentence are captured. Word embeddings were used in our project to explore associations between words and to choose possible non preferred terms on the thesaurus, even if these terms were not as frequent in the raw count.

3.0 The method

The methodologies applied in our study involve both automatic/machine based as well as manual/intellectual processes, in a collaborative setting of technical and domain experts. As a first step, the ~17,000 questions across the 109 questionnaires were tokenized and lemmatized. Abbreviations of words were resolved and stop words (e.g. articles, prepositions, etc.) removed. The non-concepts, i.e., words related to syntax, morphology, semasiology and onomasiology used to build the questions were identified and removed. The remaining "cultural heritage" words were extracted automatically using Python scripts and ranked according to their frequency across all questions (min.=1; max=609). This has yielded a total of 88,883 distinct terms, which we refer to as concepts. As put before, from all the basic subjects pertinent to the collection, we have chosen two for the proof of concept: technology (Technologie) and food (Essen), in which two different sets of domain expert in the team were involved. Technology related terms were identified manually among the most frequent by the domain experts, involving two rounds of evaluation and agreement processes, yielding a total of 186 terms. As we are dealing with a non-standard language collection, automatic identification of technology related terms would have only been partially successful, in spite of the availability of linguistic Knowledge Organization Systems as the german Wordnet (Germanet) and German Thesauri. Then, both the domain and the technical experts determined a suitable set of facets for the thesaurus, based on the concepts that were harvested and the guidelines from examples made by the CRG. The chosen facets were: trade/crafts (Gewerbe); artifacts (Artefakt), processes (Prozess), roles (Rollen) places of application (Anwendungsort), areas of application (Anwendungsbereich) and quality (Eigenschaften). Subsequently, the 186 technology related terms were assigned to these facets by the domain experts, again involving two rounds of evaluation and agreement. Terms that could not be clearly assigned, or when no agreement was reached among the domain experts were temporarily excluded (n=11). For these terms further evaluation is needed in future developments. Finally, a total of 175 technology related terms were distributed as follows: trade/crafts (n=8); artifacts (n=100), processes (n=46), roles (n=4), fields of application (n=9) and places of application (n=8). The food related terms are still being in the process of being harvested and the process has not been completed yet.

---

6 https://github.com/acdh-oeaw/exploreAT-Concepts
The chosen tools for the management of the vocabularies and data, and the display of the thesaurus hierarchies was the free and open source tool Tematres with the Visual Vocabulary addin installed. The database with the questions was kept as a Pandas dataframe, served by a Pandas REST API bridging the access to the Django REST Framework.

After all the technology related terms were chosen, the preferred and non preferred terms were added to the online thesaurus tool according to the facets in the hierarchies crafted. After all the terms were registered, the connections among the related terms were assigned. The final step was establishing links between each preferred term to the database of questions and answers that contained the term. This was made via the Tematres Multilingual Vocabularies resource (“Relations between vocabularies: RelatedMatch”), that takes the URL and generates a web link in the thesaurus management tool to the dataframe/database kept in the Pandas structure. The main advantage of this architecture is that it is comprised of an autonomous single Docker container that has the LAMP stack (Linux, Apache, MySQL, PHP/Python) tools, besides the Tematres with its enhancements, and it can be deployed quickly in any new environment. We plan to make a Docker image available - devoid of data - in the DockerHub as the initial (and outdated) version of Tematres that was used as the initial Image was updated. This will help accelerate the deployment of solutions like this in the future.

4.0 Results

In this section we will show some illustrations of the current state of the solution. Figure 1a below depicts the homepage with both the alphabetic and systematic displays for the thesaurus. The users can also make queries using the query box provided by the Tematres tool.

Figure 1b illustrates a detail on the specific basic subject of technology, showing broader and specific terms and Figure 2 presents the details for the term “Schiff” (ship) under the facet “Artifakts”. We have used bibliographic notes to provide hyperlinks (related match) to the database sources. This kind of coupling between the concept navigation tool and the data can be changed by simply substituting the URIs. In a future version, we plan on implementing a more smooth interface as we can see in the Getty Art & Architecture Thesaurus Hierarchy Display. The main advantage of the current architecture of the solution is that it allows for rapid deployment and uses free open source tools.

---

7 https://www.vocabularyserver.com/
8 https://github.com/tematres/visualVocabulary [last access: 12.12.2019]
11 https://www.djangorestframework.org/ [last access: 12.12.2019]
13 https://hub.docker.com/ [last access: 12.12.2019]
14 https://hub.docker.com/r/systemsector/tematres [last access: 12.12.2019]
Finally, we also offer a hyperbolic geometry navigation interface, provided by the Tematres solution as an optional addin (Figure 3).
5.0 Conclusions and future work

In this paper we have presented a case-study on the development of a knowledge organisation system: a facet thesaurus, for transversal knowledge discovery within a non-standard language legacy dataset. We have demonstrated that a facet approach combined with both manual (human) and automatic (machine) generated concepts are essential for eliciting cultural knowledge. This setting, combined with the online thesaurus and the connection to the database with the original concepts, have provided an effective and intuitive way for navigating and retrieving the information contained in the DBÖ resource in a structured and accessible way. Our approach makes it possible to access, link and visualise thematic knowledge also transversally, which has often proven challenging. As a next step, and as a future extension of the basic subjects being identified. Food is close to completion, and then customs, religious festivities, food, traditional medicine, professions and songs are to follow.

References
Collection Questionnaires on the example of exploreAT!. Budapest: Centre for Digital Humanities, Éötvös Loránd University, 13-15.


Indigenous Community Driven Knowledge Organization at the Interface
The Case of the Inuvialuit Digital Library

Abstract:
The goal of knowledge organization is to address the ways in which information and knowledge are conveyed, communicated, and understood. Beghtol (2002a; 2002b) argues for the principles of 'cultural warrant' and 'cultural hospitality', which make reference to the importance of taking into account cultural contexts, dimensions, and differences in knowledge organization. This notion of cultural relevance is also critical when developing digital libraries - online platforms for organizing, sharing, and providing access to resources in digital form - by, with, and for communities. The challenge of cultural relevance in digital libraries is particularly strong when working with Indigenous communities, as most digital library platforms are based on western approaches to knowledge organization. The Inuvialuit Digital Library (https://inuvialuitdigitallibrary.ca/) was developed as part of the Digital Library North project, a four-year collaboration between the Inuvialuit Cultural Centre Pitquhiit-Pitqusiit (ICC) and communities within the Inuvialuit Settlement Region (ISR) in northwestern Canada, and researchers at the University of Alberta (Edmonton, Alberta, Canada), to develop a digital library infrastructure to support access to cultural resources. Using culturally appropriate methods, the team used an iterative development process to enact a culturally reflective and responsive metadata and knowledge organization framework for the Digital Library. This community driven framework allows the Inuvialuit to tell their own story in their own words, and enhances community engagement with their Digital Library.

1.0 Introduction
The goal of knowledge organization is to address the ways in which information and knowledge are conveyed, communicated, and understood. Cultural aspects of knowledge organization and the principle of 'cultural warrant' and 'cultural hospitality' as argued by Beghtol (2002a; 2002b) make concrete references to the importance of taking into account cultural contexts, dimensions, and differences in knowledge organization. The same argument holds true about conceptualizing knowledge organization as understanding and effective communication.

Digital libraries are online environments for organizing, sharing, and providing access to resources in digital form (Borgman 1999). They are understood to be developed by, with, and for user communities. Indeed, Pang (2012) notes that “one cannot fathom a digital library without considering the social interactions driving its development, sustainability and use” (86). Ideally, then, their content, functionality, as well as metadata and knowledge organization should reflect the needs, interests, and contexts of the communities from which they originate. Baca (2003) emphasizes this when she notes that it is not enough to use some metadata standard; a metadata standard appropriate to the materials in hand and in particular the intended end-users must be selected.

Hudon (1997) and Zeng and Chan (2004) remind us of the importance of cross-cultural and cross-lingual aspects of the development of knowledge organization systems and point to the importance of cultural relevance. Boast, Bravo, and Srinivasan (2007), Clarke (2002), Srinivasan (2002) and others argue for community specific metadata and knowledge organization based on the socially constructed and contextual
nature of knowing and understanding the world. Srinivasan (2012; 2017) labels this approach “fluid ontologies”, a knowledge organization framework that is used to design a digital library in a locally appropriate manner, evolving and changing along with the community.

The development of digital libraries with Indigenous communities is particularly challenged by the fact that the technical platforms commonly used for developing them are, at their core, based on a western approach to knowledge organization (Christie 2004; 2005). Nakata (1997; 2002; 2007) explains that the digital environment is a space where Indigenous and non-Indigenous knowledge systems come into contact and where there is often tension as interactions are negotiated. He argues, however, that if Indigenous peoples are effectively and actively involved in the development and definition of the knowledge organization underlying a given system, then the power and promise of digital platforms to meet the needs and interests of the community can be achieved. While Indigenous community driven knowledge organization systems and practices have been utilized in several cultural heritage digital libraries in Australia (Bow, Christie, and Devlin 2015), New Zealand (Lilley 2015), and North America (Holland and Smith 2000), little research exists on knowledge organization practices of the Inuit in western Canada and how they can inform the development of knowledge organization systems and digital libraries (Farnel et al. 2017; Hennessy et al. 2013). The goal of this participatory, community based study is to collaboratively develop a culturally appropriate metadata and knowledge organization framework for the Inuvialuit Digital Library of cultural resources (Digital Library North 2017).

2.0 Background

The Inuvialuit Digital Library (https://inuvialuitdigitallibrary.ca/) was initially developed as part of the Digital Library North (DLN) project, a four-year collaboration between the Inuvialuit Cultural Centre Pitquhiit-Pitquisit (ICC) and communities within the Inuvialuit Settlement Region (ISR) in northwestern Canada, and researchers at the University of Alberta (Edmonton, Alberta, Canada), to develop a digital library infrastructure to support access to cultural resources. The research was contextualized in six key areas, one of which was the development of a comprehensive, culturally aware and appropriate metadata and knowledge organization framework.

To provide some context, the ISR (Figure 1) was designated in 1984 in the Final Agreement between the Inuvialuit and the Government of Canada.
The region covers approximately 91,000 km² in the western Arctic region of what is now Canada. The six communities in the region are Aklavik, Inuvik, Paulatuk, Sachs Harbour, Tuktoyaktuk, and Ulukhaktok. The population is roughly 6,500, with more than half (3,400) located in Inuvik. The language of the Inuvialuit is collectively known as Inuvialuktun, which comprises three related languages: Sallirmiutun, Uummarmiutun, and Kangiryuarmiutun (Inuvialuit Regional Corporation, 2017). While “the region has an immensely rich culture and history, its geographic remoteness poses challenges for enabling easy access to cultural heritage resources” (Farnel et al. 2016, 3). The overarching goal was to help alleviate this challenge through the development of a digital library.

3.0 Methods

Developing knowledge organization systems for Indigenous cultural heritage digital libraries requires multidisciplinary theoretical and methodological frameworks that take into account the cultural nuances of Indigenous knowledge creation, sharing, and dissemination. The definition and application of the metadata and knowledge organization framework for the Digital Library was an iterative process that incorporated culturally appropriate methods and made use of a number of information sources. A first source of information was the proposed content of the digital library itself. A second source of information was a series of interviews with the staff at the ICC who are the stewards of these resources. A third source of information was a cross-cultural, cross-disciplinary, cross-national review of the academic and professional literature to understand what had been tried and had been shown to be successful and applicable in related projects with Indigenous communities. A fourth source of information was a review of existing digital library platforms to understand their strengths and weaknesses in this given context. The fifth and final source of information was the community itself. Information was gathered from a cross-section of the community, including elders and youth, language and culture instructors, and members.
of the community at large. Information was gathered through means appropriate to this community context, including formal interviews and surveys, demonstrations and open houses, informal and targeted conversations, and user testing and usability sessions.

The information gathered through these different activities was analyzed in order to derive dominant categories and themes. Coding was both deductive and inductive. Deductive coding reflected the main research areas of the project; inductive coding reflected the themes emerging from the information itself. Categories and themes were incorporated into the metadata and knowledge organization of the Digital Library, and tested and assessed by the community. Revisions and changes were made based on community feedback, and then tested again. There was a continuous feedback loop to ensure the framework was reflective of community interests and needs as they change and evolve over time.

4.0 Examples

Cultural constructs such as language and dialect, the importance of place and resources associated with it, as well as visual interfaces have been identified as important to the ISR communities. The community driven and culturally relevant metadata and knowledge organization framework that underlies the Inuvialuit Digital Library can be seen in specific examples of metadata use and display, interface design, and content organization. The following examples highlight some key aspects of the framework as it has evolved to date.

A key message from the community has been the importance of one or more means of browsing the Digital Library. In fact, a slight preference for browse over search has been indicated. The ways in which users can explore the collection have therefore been a topic of much discussion throughout the development process. Early discussions resulted in a home page that allowed for browsing by type and collection, as well as by featured images or exhibits.

Over time it became clearer that there were certain key pathways into the Library that the community would like to see privileged. Figure 2 shows the current version of the home page which has the most important pathways, such as places and language resources, more prominent, with the additional pathways of resource type and a featured image still available.
Given that the ICC’s mandate is language and culture revitalization, language learning resources, many developed by the Centre itself, represent a substantial portion of the overall Library and are seen as critical resources for highlighting. Initially, language resources were simply noted as one type of collection, accessible through browse by collection or by type, and the landing page had a simple structure of labelled images for each language (Figure 3).

However, as discussions continued our community collaborators discussed how they would like to do more with these collections. Not only did they want them to be a more prominent pathway into the Library, they also wanted to add additional contextual information to the landing page. The current version of the language resources landing page (Figure 4) is much richer in content as it includes information about each language and shows a map of the areas in the region where each is spoken, and clicking on any of the language names takes the user to a listing of all the resources dealing with that language.
The Inuvialuit, like all Indigenous peoples, have a strong connection to land and place. For this reason, there has been strong interest in the ability to browse the Digital Library content by place. An early version of this functionality allows the user to find a single item on a map (Figure 5) and click through to view it, and from there use the metadata in the record to browse other items associated with that same place.

This functionality is well liked and is still available in the Library. However, this was not quite what the community had in mind. What has been described is the ability to start your browsing with a map, and to narrow into specific places and find all items associated with it. Figure 6 shows an interim version of this which is currently part of the Library. At the moment, this map includes only the six community names. If a user clicks on a community name (e.g., Ulukhaktok), they will be taken to a set of items with that place in the metadata. The Inuvialuit Regional Corporation (IRC), the ICC’s parent body, is currently working on a traditional place names map which will be much, much
richer than the one currently in use. The plan, once this map is complete, is to use it in place of the one currently in use in the Digital Library.

Figure 6. Current map based browse and results from clicking on Ulukhaktok on the map

An important component of a knowledge organization framework is the choice of metadata elements used to describe the resources, as well as what those elements are called, and what they contain. This is no different in the case of the Inuvialuit Digital Library, and developments in this area have been a large part of the community driven process.

With many of the resources in the Library having a linguistic aspect, the ability to capture Language and Dialect, as well as Original Dialect in cases where there was a translation, was identified as critical from the earliest days of the project. The ability to see this information immediately for any resource resulted in its prominent place on any item screen, and the desire to make it easy to find other resources with the same language or dialect prompted the metadata element to be made browsable.

With the cultural importance of place and land, family and community, the ability to capture in the metadata the places and people associated with a resource is critical. Early input from the community led to the renaming of elements to make them more relevant and usable; Creator and Contributor were combined into a single element and renamed People; Spatial Coverage was renamed Places. The Inuvialuit, like all Indigenous peoples in Canada, are victim to ongoing efforts to erase their culture, heritage, and language, including traditional names for people and places. These names and the traditions around them are being reclaimed by the community, and so including them in the metadata for resources in the Digital Library is extremely important. However, we have also heard of the importance of retaining in some way the colonial forms as well, as there are still in use and do represent an important part of the history of the Inuvialuit. And so a balance is struck, with the metadata including both but privileging the traditional. And the framework is also flexible enough to account for alternative spellings and dialect variations. Figure 7 shows these various aspects of resource description.
A further aspect of the metadata description to highlight deals with the ways in which the subject matter of the resource is described. The community recognizes the value of using existing vocabularies and term lists for making the Library usable and sustainable. But there is also strong interest in being able to use the local language where and when it makes sense, and to allow for spelling and dialect variations as well. And so a growing local list of such terms is in use in the Digital Library, as can be seen in the description in Figure 8, which includes the local English and Inuvialuktun terms for parka.
5.0 Conclusion

The most powerful experiences with digital collections occur when the knowledge structure and architecture are harnessed to the interests and needs of the community. The metadata and knowledge organization framework for the Inuvialuit Digital Library, and the collaborative methods used to develop it, demonstrate that knowledge organization is communication, understanding, and development. Noted Maori scholar Linda Tuhiwai-Smith reminds us that “the collective memory of imperialism has been perpetuated through the ways in which knowledge about Indigenous peoples was collected, classified and then represented in various ways … through the eyes of the West, back to those who have been colonized” (2012, 31). A community driven metadata and knowledge organization framework enables the Inuvialuit to tell the story they want to tell, in the way they want to tell it, pushing back against the story being told by others, re-centering the community and putting control back where it belongs.

References


Towards an Open, Inclusive and Sustainable Knowledge Organization Models

Abstract:
In an increasingly globalized context, multilingualism and multiculturalism have become major preoccupations for Knowledge Organization (KO) which have to be as fair as possible to ensure and sustain knowledge organization as a driver for development. Indeed, over time, the gap between languages of dominant nations or civilizations and other languages has been growing. In this research, we describe, evaluate and present first results of our sustainable and open access knowledge organization model. The model is based on a paradigms that permit different types of contributors, including volunteers as well as scientific and scholarly communities from across borders, languages, nations, continents, and disciplines to take part in the knowledge organization process in an efficient and dynamic way. Recent experiments with this model have been conducted on transnational literary texts as well as in the arena of crowdsourcing cultural heritage knowledge and collections enrichment.

1.0 Introduction

The impact of the digital revolution on the preservation, organization, and sharing of human knowledge encoded by languages constitutes an extraordinarily rich phenomenon, characterized by both productive opportunities as well as obstacles and threats. In the first instance, digital has created tremendous opportunities in terms of accessing knowledge. New technologies also constitute a step forward in terms of public inclusion and awareness. In fact, the general public can be included and integrated, thanks to social networks and collaborative platforms, to provide mass dissemination of human knowledge. Nevertheless, there are numerous barriers that prevent sustained knowledge diversity as described by Hudon (1997), Beghtol (2002), Fraisse et al. (2019), and Barát (2008). Language is the most important barrier; as language diversity is decreasing, the preservation and transmission of such knowledge is at risk. The ever growing scientific and political interests in making knowledge open, accessible and sustainable has sparked major interest in many parts of the scientific community. Some disciplines have been concerned with problems of knowledge dissemination for a long time. Library and Information Science (LIS) is such a discipline. As a gateway to knowledge and culture, the field of LIS holds a long history on collecting, storing, organizing, and sharing access to knowledge as described by the pioneer of Documentation Studies Paul Otlet (1934). To this purpose, Knowledge Organization Systems (KOS), Information Retrieval Systems (IRS) and metadata exchange standards, among others, have been developed to meet the opportunities arising through the development of new technologies. Collections of the world's great libraries have been made available to the public through large-scale digitization. The Online Computer Library Center (OCLC), dedicated to the public purpose of furthering access to the world's information, produces and maintains WorldCat, the largest online public access catalog (OPAC) in the world. WorldCat itemizes the collections of 72,000 libraries in...
170 countries and territories. Multilingual online digital libraries and archival projects collect documents and make them available to a wide audience.

2.0 The role of library and information science in building a global, shared knowledge community

More than a century ago, Paul Otlet, the pioneer of Documentation Studies, envisioned a universal compilation of knowledge and the technology to make it globally available. He wrote numerous essays on how to collect and organize the world's knowledge (Otlet 1934). The ever-growing number of digital documents and scientific and political interests in making them openly available all over the world has led to the creation of new digital collections in a broad range of fields and languages. Several Registries of Open Access Repositories (ROARs) hosted by national and international organizations and universities, have been developed. For example, The Library of Congress\(^1\) has digitized approximately 164 million items in virtually all formats, languages, subjects, and periods. These collections are broad in scope, including research materials in more than 470 languages and multiple media. The Europeana collection\(^2\), launched in 2008 and funded by the European Commission, contains over fifteen million digitized paintings, drawings, maps, photos, books, newspapers, letters, diaries, etc., from fifteen hundred institutions. However, the language barrier is a key issue that Knowledge Organization Systems (KOS) have to address as described by Hudon (1997; 1998) and Agnes Hajdu Barat (2008). Indeed, over time, the gap between languages of dominant nations or civilizations and other languages has been growing. Although KOS include knowledge encoded in under-resourced languages, their use and exploration is still limited.

3.0 Current situation

3.1 Crowdsourcing as a means of decentering institutional authority and expanding the representation of different languages and cultures

Since the year 2000 online crowdsourcing projects have proliferated in science, humanities, and cultural heritage fields. Hundreds of cultural heritage institutions have embarked on these projects, many of which explicitly invite people from diverse walks of life to transcribe, annotate or highlight text, speech, typed or handwritten documents. A wide spectrum of languages, historical periods, materials, and geographic areas are represented by these projects and the people who participate in them as described by Van Hyning (2019) and Ridge (2014). As more transcriptions and tags become available in different languages, cultural heritage institutions better represent the peoples and cultures in their collections, and the patrons and communities they serve. By inviting volunteers in to the process of transcribing, translating, and tagging, institutions have the opportunity to co-create new knowledge, and make new discovery pathways through collections. These relationships can ultimately decenter traditional power dynamics and concepts of authority in knowledge systems, often for the better, though as Eveleigh (2014) demonstrates, the breakdown of barriers between professional practice and the knowledge of external participants is not inevitable with all crowdsourcing projects, but

\(^{1}\) https://www.loc.gov
\(^{2}\) https://www.europeana.eu
rather requires careful project design and strategies of volunteer engagement. Examples of under-represented language crowdsourcing projects include the City Archive of Leuven project to transcribe more than 950,000 Dutch-language register pages from the Leuven court of Aldermen during the years 1362 to 1795; the Ancient Lives project, which launched on the Zooniverse.org platform in 2011, described by Williams et al. (2014), which asked online volunteers to transcribe fragments of ancient Greek texts from papyri fragments; and the Rediscovering Indigenous Languages project crowdsourced the transcription of historic word lists, records and other documents relating to indigenous Australian languages. Many of these communities are self-sustaining, self-organizing, and productive of new knowledge, as well as decentralized—participants can translate and describe code and systems, but they can also contribute new functionality to them. This type of community co-creation might serve as a fruitful model for cultural heritage crowdsourcing, in which authority and the creation of descriptive records is still often overwhelmingly concentrated within institutions, rather than shared with the communities that originate cultural artefacts, texts, music, dance, and other outputs.

3.2 Increasing demand and need for global knowledge sharing and access

According to the Sapient Globalization Report there are over 6,700 living languages in the world; the fifteen most popular languages are spoken by 49.5% of the world’s population, while the other 51.5% of the world’s population speak 6,600 languages. Yet, only about 6% of the world’s population speak English. Of the world’s 6,000+ languages only a small fraction, a dozen or so, currently enjoy the benefits of modern information technologies and knowledge organization systems. A larger but still modest number, close to a hundred, have the so-called Basic LAnguage Resource Kit (BLARK): monolingual and bilingual corpora, machine readable dictionaries, terminologies, thesauri, ontologies and the like as described by Steven Krauwer (2003) and Antti Arppe et al. (2016). Preserving knowledge diversity and ensuring the right of all people to access knowledge in their mother tongue is the main goal of the Information for All Programme (IFAP) created by UNESCO. Several research work have called for cultural and linguistic diversity as described by e.g. Adler et al. (2016), Beghtol (2005), Dahlberg (1992), López-Huertas (2016), and Mustafa El Hadi (2015). In a previous research work Beghtol (1986; 2001) introduce the concept of cultural warrant. Fisher Fishkin (2011) introduced and described a new model for data curation and sharing by inviting colleagues around the world to collaborate on Digital Palimpsest Mapping Projects (DPMPs), or “Deep Maps”. Deep Maps, curated collaboratively by scholars in multiple locations, would put multilingual digital archives around the globe in conversation with one another, using maps as the gateway.

---
3 https://www.itineranova.be/in/home
4 https://transcripts.sl.nsw.gov.au

© Ergon - ein Verlag in der Nomos Verlagsgesellschaft
4.0 The open, inclusive and sustainable knowledge organization models
4.1 Basic principles
4.1.1 From closed, discontinuous and out of context to open, continuous and in context knowledge organization models

Our solution aims to move from a closed, discontinuous, and out of context to open, continuous, and in context knowledge organization model. The basic concept is based upon the software localization paradigm proposed by Fraisse (2010) and Fraisse et al. (2009) and promoting the right of all people to use software in their mother tongue. It consists of renouncing the idea of perfect and complete knowledge and publishing partial knowledge with variable quality, which will be improved incrementally during the use of the knowledge organization system. Therefore, the knowledge organization process will be ongoing and improve continuously. The new process permits the incremental augmentation of both quality and quantity. The best known example of this is the Wikipedia community, in which knowledge is added and improved continuously by contributors.

4.1.2 From exclusive, unilateral and unsustainable to inclusive, collaborative and sustainable

Although, current knowledge organization models seems impossible for most languages and even less so for endangered ones, both for reasons of cost, and quite often a scarcity or even lack of expert in these languages. Our solution aims at involving non-experts such as volunteer contributors and especially end-users. These groups have the capacity to participate effectively, since they have a better knowledge of the target language (generally their native language) as well as the context of knowledge being processed.

4.2 The ROSETTA knowledge organization model

An early implementation of the open, inclusive and sustainable knowledge organization model described above was implemented and experimented under the international research project ROSETTA 5 funded by the France-Stanford center for interdisciplinary studies. The main goal of this project consists of defining a knowledge organization model for translated literary texts as well as related scientific documents to these translations. As described by Fraisse et al. (2019), the proposed model is open, inclusive and sustainable. It is based on contributions of end-users as well as scientific and scholarly communities from across borders, languages, nations, continents, and disciplines. It consists in collecting knowledge about all worldwide translations of one original work and sharing that data through a digital and interactive global knowledge map. The proposed sustainable model allows different types of volunteers and contributors to take a part in the knowledge organization process in an efficient, dynamic and symbiotic way: while using the knowledge map, volunteers and contributors who know the local culture and language can participate by adding missing information about a given translation of an original work. Volunteers and contributors could be scholars or simply citizens interested in preserving knowledge diversity. We define global

---

5 https://francestanford.stanford.edu/projects/rosetta-resources-endangered-languages-through-translated-texts
knowledge $GK$ about an original work $ow$ as a set of knowledge $K$ about different translations $t$ of $ow$:

$$GK_{ow} = \{K_{ow}^{t_1}, K_{ow}^{t_2}, ..., K_{ow}^{t_n}\}$$

where a knowledge $K_{ow}^{t_i}$ about a given translation $t_i$ is a set of key properties as:

$$K_{ow}^{t_i} = \{\text{title}, \text{target language}, \text{translator name}, \text{publication date}, \text{full text}\}$$

Related knowledge was organized collaboratively and in context by scholars through an interactive and online global knowledge map. As described in Figure 1, the map displays all knowledge about all existing documents related to a given literary work. Each document is represented by a node on the world map, which could be considered as “completed” when all required knowledge is provided and “partially completed” when it lacks some knowledge. Nodes are updated incrementally by transnational end-users and scholars through the map. Indeed, during the map exploration, the end-user could edit any node to add missing knowledge.

Figure 1: The global knowledge map representing existing translations of Adventures of Huckleberry Finn. The bubble over Brazil is highlighted, displaying the relevant information for the Portuguese translations from Brazil.

4.3 Co-creating and crowdsourcing knowledge of folklife and music traditions through the Library of Congress

Traditions of collaborative knowledge creation in cultural heritage are perhaps rarer than they should be, but there are precedents in this sector as well. The twentieth-century folklorist Alan Lomax devoted his life to recording, celebrating, and promoting folk artists and tradition bearers in America, the Caribbean, and Europe. He conducted extensive fieldwork trips during which he produced audio recordings and extensive notes about the people he met, and their traditional arts. His goal was to demonstrate the value of traditional arts, and challenge what he saw as a hegemonic media and cultural system in America and Europe which failed to make room for cultural differences and killed off diversity. As Harvey et al. (2017) argue, Lomax was critical of “a centralized mediascape through which was broadcast an industrial American monoculture”. “Too few transmitters and too many receivers” was his central complaint. He was frustrated with the myopic unilateralism of corporate programming, which he saw operating through an “over-powerful, over-rich, over-reaching” communication system. His
answer to this was what he termed “cultural equity”: the right for folk communities—what he called “little bubbles of song and delight and ways of life and cookery,” encompassing “hundreds of thousands of these little generators of the original” - to have their voices heard and their traditions represented.” Lomax ultimately recorded over 1000 cultural groups, and hundreds of under-represented languages. He established the Association for Cultural Equity to advocate for folk artists, and donated his field notebooks, recordings, letters, and other papers to the Library of Congress where he helped to establish the American Folklife Center (AFC). In 2015, the AFC digitized Lomax’s papers and made them available online. In 2019, AFC partnered with a new crowdsourcing effort called By the People® at the Library of Congress, to crowdsource the transcription, review, and tagging of these papers. By the People’s goals are to engage a diverse volunteer base with cultural heritage preserved at the Library of Congress; to generate transcriptions that will improve online search at the document level, and to provide transcriptions that can be read by screen readers, in order to assist people with visual or cognitive impairments, and those who can’t read original handwriting. By the People launched in October 2018 and to date volunteers have transcribed over 100,000 pages from a variety of collections including the papers of Rosa Parks, Walt Whitman, President Abraham Lincoln, and leading suffragists such as Susan B. Anthony and Mary Church Terrell. Volunteers are encouraged through the site itself, emails, in-person events, and social media to explore the documents, ask questions, speak with one another, and Library employees about their findings, struggles, joys, and what they’re learning. Their knowledge is taken back into the Library website in the form of transcriptions and enhanced metadata.

By the People is a natural extension of Alan Lomax’s efforts to build “two-way bridges” and “two-way inter-communication systems’ for traditions presented in any medium” as described by Baron (2012). Documents in “The Man Who Recorded the World: On the Road with Alan Lomax” by the People transcription Campaign include materials in Haitian Creole, and dialects of Swedish, Polish, Danish, Hungarian, and other languages spoken by nineteenth- and twentieth-century migrants to the American Midwest, which volunteers transcribe in the original language. In addition to reaching out to over 30,000 registered volunteers to encourage them to participate in the project, AFC folklorists reached out to several descendants of the tradition bearers whom Lomax originally recorded to encourage them to contribute to By the People, and bring their knowledge to bear in this next phase of folklife preservation and exploration.

4.4 Crowdsourcing multilingual knowledge: The Zooniverse platform

One current example of Knowledge Organization principles applied through crowdsourcing is Scribes of the Cairo Geniza®, a collaboration between Zooniverse®, the University of Pennsylvania Libraries, and more than half a dozen research institutions who have provided digital geniza images for the project®. The Zooniverse is the largest platform in the world for online crowdsourced research, with more than 250 projects launched since its inception in 2009. As of writing, the Zooniverse has more

---

6 https://crowd.loc.gov
7 https://www.scribesofthecairogeniza.org
8 https://www.zooniverse.org
9 https://www.scribesofthecairogeniza.org/about#provenance
than 1.9 million registered volunteers, who have collectively produced over 460 million classifications on crowdsourcing projects from a variety of disciplines, including astronomy, biology, ecology, climate science, history, and social science as explained in Blickhan et al. (2019). In 2015, Zooniverse launched the Project Builder\(^\text{10}\), a tool which allows anyone to create and run their own crowdsourcing project, hosted on Zooniverse, free of charge. The platform is maintained by teams based at the University of Oxford (Oxford, UK), the Adler Planetarium (Chicago, IL), and the University of Minnesota Twin Cities (Minneapolis, MN). While the majority of users thus far have been from English-speaking countries, the Zooniverse community is international. To reflect the global userbase, the Zooniverse team created a translation interface, which allows projects to be translated either by project team members, or by volunteers who want to help make a project available for a specific community of speakers. Along with the relatively recent option of a multilingual interface, the Zooniverse has featured multilingual project content since its early days. Ancient Lives\(^\text{11}\), which launched in 2011, invited volunteers to transcribe fragments of the Oxyrhynchus papyri. In order to open up participation to members of the public who were not fluent in Ancient Greek, the team created a clickable keyboards that volunteers could use to transcribe fragments through character matching. The Scribes of the Cairo Geniza project launched a transcription interface in 2019 with a similar type of assistive keyboard, as well as a multilingual user interface in Arabic, English, and Hebrew. In Scribes of the Cairo Geniza, volunteers are asked to help sort and transcribe fragments of the Cairo Geniza, a corpus of discarded fragments of pre-modern manuscripts discovered in the Ben Ezra synagogue in Fustat (now known as Cairo). The project is broken down into a series of workflows. The Sorting workflow asks volunteers to classify fragments as being written in either Arabic script, Hebrew script, or both. Based on the script type identified, volunteers are then asked to identify specific visual features like page layout, evidence of binding, etc., which can be added to each fragment’s record and assist in the identification process. Once a fragment has been sorted, it is sent to one of four transcription workflows: Easy or Difficult Arabic, or Easy/Difficult Hebrew. Separating workflows based on task type allows volunteers to choose how they wish to contribute based on their comfort level with the different tasks. For example, volunteers who are unable to read Hebrew or Arabic script are able to participate in the Sorting workflow, which offers an introductory tutorial as well as resources on identifying the differences between Hebrew and Arabic script. Volunteers who are fluent in Hebrew and/or Arabic, or who are able to confidently read either script, may choose to participate in the various transcription workflows. The transcription workflows feature clickable keyboards, which function as an additional linguistic and paleographic resource for transcribers who may need additional visual cues to aid in the transcription process.

5.0 Conclusion
Knowledge Organization is facing a range of highly challenging issues considering the diversity of knowledge encoded in different languages and in particular those encoded in vulnerable and under-resourced ones. In this paper we described and explored an open, inclusive and sustainable knowledge model that permit different types

\(^{10}\) https://www.zooniverse.org/lab
\(^{11}\) https://www.ancientlives.org
of contributors, including volunteers as well as scientific and scholarly communities from across borders, languages, nations, continents, and disciplines to take part in the knowledge organization process in an efficient and dynamic way. We explored examples of modern online crowdsourcing, as well as some of the historic attitudes within cultural heritage institutions that have led to or stood in contrast to ideas of co-production or collaboration between institutional gatekeepers and patrons of diverse cultural backgrounds. Crowdsourcing has huge potential to expand the representation of vulnerable languages and cultural practices within the cultural heritage record, and to radically expand the base of people who contribute to the knowledge that is preserved and treated as authoritative by cultural heritage organizations, academia, and other domains.

References


Samantha Blickhan, Coleman Krawczyk, Daniel Hanson, Amy Boyer, Andrea Simenstad, Victoria Hyning, and Victoria van Hyning. 2019. “Individual vs. Collaborative Methods of Crowdsourced Transcription.” Journal of Data Mining and Digital Humanities: hal-02280013v2. https://hal.archives-ouvertes.fr/hal-02280013v2/document


Hudon, Michèle. 1997. “Multilingual Thesaurus Construction-Integrating the Views of Different Cultures in One Gateway to Knowledge and Concepts”. In: Information Services and Use 17: 11–123.


Williams, Alex C., John F. Wallin, Haoyu Yu, Marco Perale, Hyrum D. Carroll, Anne-Francoise Lamblin, Lucy Fortson, Dirk Obbink, Chris J.Lintott, and James H. Brusuelas. 2014. “A Computational Pipeline For Crowdsourced Transcriptions Of Ancient Greek Papyrus Fragments”. In 2014 IEEE International Conference on Big Data (Big Data), Washington, DC, 100-105. doi: 10.1109/BigData.2014.7004460.
New Formats, Shifting Fortunes
Late-Twentieth-Century KO in the Wild

Abstract:
Three candidates for the knowledge organization (KO) systems that enjoyed the widest, most popular usage in the late twentieth century are (1) Encyclopædia Britannica’s Propædia or “Outline of Knowledge,” (2) the “Synopsis of Categories” at the heart of HarperCollins’ Roget’s International Thesaurus, and (3) OCLC’s Dewey Decimal Classification. Surprisingly, given the popularity of these systems (which continues in the latter two cases into the 2020s), only the last has received special attention in the ISKO community. The goal of this paper is to compare the function, form, and content of each of the three systems, in the context of a taxonomy of evaluation methods for KO applications that takes into account similarities and differences in formats and purposes.

1.0 Introduction
Classification involves the identification of “groups of things [that] can then be combined and arranged to make a . . . system” (Beghtol 2010, 1045). In the remainder of this paper, I compare and contrast three cases in which classification has been carried out in the production of a practical system of a different kind. Each system is distinguished by the particular kind of things that make up its groups.

In the first case, the system is a scheme for describing and arranging the subjects (i.e., topics) of entries in a general English-language encyclopedia (and thus for describing and arranging those entries themselves). An encyclopedia is “a literary work containing extensive information on all branches of knowledge, usually arranged in alphabetical order” (OED Online, December 2019).¹ The encyclopedia in the case in question is the 15th edition of the Encyclopædia Britannica (EB; Hoiberg 2010);² the scheme is the “Outline of Knowledge” (OoK), commonly known as the Propædia.

In the second case, the system is a scheme for describing and arranging the meanings of entries (i.e., concepts) in a general English-language thesaurus (and thus for describing and arranging those entries themselves). A thesaurus is “a collection of concepts or words arranged according to sense” (OED Online, December 2019).³ The thesaurus in the case in question is the 8th edition of Roget’s International Thesaurus (RIT; Kipfer 2019);⁴ the scheme is the “Synopsis of Categories” (SoC).

In the third case, the system is a scheme for describing and arranging the subjects (i.e., topics) of entries in any general English-language library catalog (and thus for describing and arranging those entries themselves, as well as the resources that are themselves described by those entries). A catalog is “usually distinguished from a mere list or enumeration, by systematic or methodical arrangement, alphabetical or other order, and often by the addition of brief particulars, descriptive, or aiding information” (OED Online, December 2019). The systematic arrangement adopted for the subjects of

¹ For the history of encyclopedias, see Loveland (2019).
² For the history of the Encyclopædia Britannica, see Whiteley (1992).
³ For the history of thesauri, see Hüllen (2009).
⁴ For the history of Roget’s Thesaurus, see Hüllen (2004).
library resources is typically set down in a library classification scheme. The scheme in the case in question is the *Dewey Decimal Classification (DDC; Mitchell 2011)*.

### 2.0 Method of Construction

Each scheme was originally the work of one pioneer, collecting names of items (fields, concepts, subjects) and grouping items into classes and subclasses manually, on the basis of individual experience and self-proclaimed expertise. The first instance of the OoK, published in 1974, revised in 1985, and left to collect dust alongside the rest of the final print edition of *EB* in 2012, was compiled by Mortimer J. Adler (1902–2001). The first instance of the SoC, published in 1852, was compiled by Peter Mark Roget (1779–1869); the version currently in use (Kipfer 2019) was developed by Robert L. Chapman (1920–2002) for the 1992 edition of *RIT*. The first instance of the *DDC*, published in 1876, was compiled by Melvil Dewey (1851–1931); much later, successive editions would be the work of small teams of editorial staff members, each led by a single editor-in-chief and supported by an international advisory board (the Editorial Policy Committee, EPC).

Each scheme is essentially subjective, in that the contents of classes, and the relationships among them, are not somehow “read off” an objective reality; classes are assigned to positions in a tree structure on the basis of an individual’s perceptions and judgments. These perceptions and judgments are bound to vary greatly in accordance with differences in personal attitudes, preferences, and goals, as well as with differences in the sociocultural contexts characteristic of different times and places. So it is to be expected that schemes developed by different people for different purposes, even if they are intended for general rather than special application, will vary in form at the macro-level, let alone at the micro-level. What is remarkable, given this expectation, is the degree to which the three schemes are in fact similar in certain aspects of their form, as well as in their function.

### 3.0 Function

The primary functions of the three schemes—threefold in each case—are similar, as demonstrated by the following summary.

The OoK is a scheme for describing, classifying, and arranging the entries in an encyclopedia according to the fields (i.e., the disciplines) to which those entries contribute; the SoC is a scheme for describing, classifying, and arranging the words and phrases in a lexicon according to the meanings of those words and phrases; the *DDC* is a scheme for describing, classifying, and arranging the resources in a collection according to the subjects (i.e., the topics) of the works that those resources instantiate. The results of applying any of the schemes to any given collection of items (entries, words, or resources) are (a) that items with similar characteristics are brought close together, as members of the same class, and (b) that classes with similar characteristics

---

5 For the (pre-1930) history of library classification, see Richardson (1930).
6 For the history of the *Dewey Decimal Classification*, see Miksa (1998).
7 Note we are not talking here of the assignment of resources, words, or entries to classes, but of the initial assignment of classes to positions that comprises the act of original creation of a scheme.
are brought close together, as proximate classes. In general, similarity among fields, meanings, or subjects is represented in the scheme by proximity among classes.

The other two primary functions of the schemes are (a) indexical and (b) pedagogical. Each scheme provides a means for readers (a) to search for, locate, and access items of interest, and (b) to learn about the world, both by interacting with those items and by studying the scheme itself. In particular, each scheme provides a sense of the shape, size, and structure of the totality of general knowledge.

4.0 Form

Each scheme takes the form of a tree whose lowest level of branches is relatively small in number: The OoK has ten main classes; the SoC has fifteen; the DDC has ten. In each case, the main classes are divided into a small number of subclasses, which are in turn divided into sub-subclasses, and so on. The DDC in its current form is the odd one out, in the sense that it is several orders of magnitude larger and more complex than the other two; but its basic structuring principle—the hierarchy—is the same. The main classes of each scheme are presented in Tables 1, 2, and 3.

5.0 Content

The “Outline of Knowledge” that formed the bulk of the EB’s Propædia is a list of 15,000 subject headings, subheadings, sub-subheadings, etc., arranged systematically in a 7-level taxonomic structure. Each part is divided into a number of divisions (42 in total), each of which is divided into a number of sections (189 in total), in which each topic covered is outlined; at the end of each section, a list is given of suggested readings in the Macropædia and Micropædia. This structure was Adler’s, in his capacity as director of planning for the new EB, and had been worked out between 1965 and 1968. The rearrangement reflected Adler’s “love for classification and bringing a unity to knowledge,” and more generally his “interest in self-education” (Whiteley 1992).

Adler was adamant that this structure should be viewed, not as a line or tree, but primarily as a circle. For Adler (1974, 6), the circle is a “powerful metaphor”: “with the circular arrangement of the parts, and with the rotation of the circle, the reader can begin anywhere in the circle of learning and go to adjacent parts around the circle; or, moving along interior transecting lines, the reader can go from any part across the circle to parts that are not adjacent on the circumference.” Moreover, the OoK’s part 10 might be placed in the center of the circle, reflecting a distinction between (a) “what we know about the world . . . by means of the various branches of learning or departments of scholarship” (parts 1 through 9) and (b) “what we know about the branches of learning or departments of scholarship—the various academic disciplines themselves” (part 10). The latter is what Quinton (1974, 9) calls “knowledge about knowledge, or knowledge of the second order”: i.e., the fields of logic, mathematics, science (“conceived as a knowledge-seeking activity, not as a set of findings”), history and the humanities, and philosophy.
Table 1. Main classes of the OoK (2010).

<table>
<thead>
<tr>
<th>#</th>
<th>Caption</th>
<th>f</th>
<th>f/\Sigma f</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Matter and Energy</td>
<td>39</td>
<td>8%</td>
</tr>
<tr>
<td>2</td>
<td>The Earth</td>
<td>25</td>
<td>5%</td>
</tr>
<tr>
<td>3</td>
<td>Life on Earth</td>
<td>44</td>
<td>9%</td>
</tr>
<tr>
<td>4</td>
<td>Human Life</td>
<td>24</td>
<td>5%</td>
</tr>
<tr>
<td>5</td>
<td>Human Society</td>
<td>46</td>
<td>10%</td>
</tr>
<tr>
<td>6</td>
<td>Art</td>
<td>39</td>
<td>8%</td>
</tr>
<tr>
<td>7</td>
<td>Technology</td>
<td>34</td>
<td>7%</td>
</tr>
<tr>
<td>8</td>
<td>Religion</td>
<td>35</td>
<td>8%</td>
</tr>
<tr>
<td>9</td>
<td>The History of Mankind</td>
<td>132</td>
<td>28%</td>
</tr>
<tr>
<td>10</td>
<td>The Branches of Knowledge</td>
<td>45</td>
<td>10%</td>
</tr>
</tbody>
</table>

\( f = \text{count of pages for each class} \)
\( \Sigma f = \text{total count of pages} = 464 \)

Table 2. Main classes of the SoC (2019).

<table>
<thead>
<tr>
<th>#</th>
<th>Caption</th>
<th>f</th>
<th>f/\Sigma f</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The Body and the Senses</td>
<td>92</td>
<td>9%</td>
</tr>
<tr>
<td>2</td>
<td>Feelings</td>
<td>65</td>
<td>6%</td>
</tr>
<tr>
<td>3</td>
<td>Place and Change of Place</td>
<td>86</td>
<td>8%</td>
</tr>
<tr>
<td>4</td>
<td>Measure and Shape</td>
<td>57</td>
<td>5%</td>
</tr>
<tr>
<td>5</td>
<td>Living Things</td>
<td>12</td>
<td>1%</td>
</tr>
<tr>
<td>6</td>
<td>Natural Phenomena</td>
<td>8</td>
<td>1%</td>
</tr>
<tr>
<td>7</td>
<td>Behavior and the Will</td>
<td>196</td>
<td>18%</td>
</tr>
<tr>
<td>8</td>
<td>Language</td>
<td>42</td>
<td>4%</td>
</tr>
<tr>
<td>9</td>
<td>Human Society and Institutions</td>
<td>77</td>
<td>7%</td>
</tr>
<tr>
<td>10</td>
<td>Values and Ideals</td>
<td>68</td>
<td>6%</td>
</tr>
<tr>
<td>11</td>
<td>Arts</td>
<td>20</td>
<td>2%</td>
</tr>
<tr>
<td>12</td>
<td>Occupations and Crafts</td>
<td>20</td>
<td>2%</td>
</tr>
<tr>
<td>13</td>
<td>Sports and Recreation</td>
<td>17</td>
<td>2%</td>
</tr>
<tr>
<td>14</td>
<td>The Mind and Ideas</td>
<td>256</td>
<td>24%</td>
</tr>
<tr>
<td>15</td>
<td>Science and Technology</td>
<td>59</td>
<td>5%</td>
</tr>
</tbody>
</table>

\( f = \text{count of categories in each class} \)
\( \Sigma f = \text{total count of categories} = 1075 \)
Table 3. Main classes of the *DDC* (2011).

<table>
<thead>
<tr>
<th>#</th>
<th>Caption in 2011</th>
<th>Caption in 1876</th>
<th>$f$</th>
<th>$f / \Sigma f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>000</td>
<td>Computer science, information &amp; general works</td>
<td>[no caption]</td>
<td>97</td>
<td>4%</td>
</tr>
<tr>
<td>100</td>
<td>Philosophy &amp; psychology</td>
<td>Philosophy</td>
<td>67</td>
<td>3%</td>
</tr>
<tr>
<td>200</td>
<td>Religion</td>
<td>Theology</td>
<td>158</td>
<td>7%</td>
</tr>
<tr>
<td>300</td>
<td>Social sciences</td>
<td>Sociology</td>
<td>602</td>
<td>25%</td>
</tr>
<tr>
<td>400</td>
<td>Language</td>
<td>Philology</td>
<td>58</td>
<td>2%</td>
</tr>
<tr>
<td>500</td>
<td>Science</td>
<td>Natural Science</td>
<td>307</td>
<td>13%</td>
</tr>
<tr>
<td>600</td>
<td>Technology</td>
<td>Useful Arts</td>
<td>527</td>
<td>22%</td>
</tr>
<tr>
<td>700</td>
<td>Arts &amp; recreation</td>
<td>Fine Arts</td>
<td>239</td>
<td>10%</td>
</tr>
<tr>
<td>800</td>
<td>Literature</td>
<td>Literature</td>
<td>91</td>
<td>4%</td>
</tr>
<tr>
<td>900</td>
<td>History &amp; geography</td>
<td>History</td>
<td>281</td>
<td>12%</td>
</tr>
</tbody>
</table>

$f = \text{count of pages for each class}$

$\Sigma f = \text{total count of pages } = 2427$

The 8th *International* edition’s 15 main classes into which its 1,075 categories of words and phrases are grouped are outlined in a “Synopsis of Categories” (Kipfer 2019, xix–xxxix), just as Peter Mark Roget’s 6 main classes were in his 1st edition of 1852. The current structure was introduced by Robert L. Chapman (emeritus professor of English, Drew University, Madison, NJ) as editor of the 5th *International* edition of 1992. Prior editions had retained Roget’s original structure with remarkably little change, as have all U.K. editions to date. Chapman should be credited for the most far-reaching of all revisions made since 1852. In his work on the *Thesaurus*, Chapman acknowledges the help of the philosopher Charles Courtney (Drew University), and the cognitive psychologist George Miller (Princeton University), well known for his work on WordNet, the lexical database of English.\(^8\)

Somewhat remarkably, the ten main classes in the *DDC* have survived into the twenty-first century in essentially the same form that they had in Melvil Dewey’s original plan of 1876. Precursors of Dewey’s scheme include Nathaniel B. Shurtleff’s *Decimal System for the Arrangement and Administration of Libraries* (1856) and William Torrey Harris’s scheme for the classification of books in the St. Louis Public School Library (1870), as well as the systems of Bacon (in *The Advancement of Learning*, 1605) and Hegel (in *Enzyklopädie der philosophischen Wissenschaften*, 1817) for the classification of the sciences, whose influences on Dewey have been much debated down the years.\(^9\) Wiegand (1998, 189), for example, concludes that Dewey chose Harris’s hierarchy for his own scheme “because it fit the Anglo-Saxon world into which he was born, a world further refined by the ... tradition, curriculum, and faculty” of the tiny Amherst College where Dewey had studied and worked. Given its widespread use around the world, “it is probably also fair to say that for the past century [the *DDC*] has quietly—almost invisibly—occupied an influential position as one of the

---

\(^{8}\) See [http://wordnet.princeton.edu/](http://wordnet.princeton.edu/).

\(^{9}\) For the history of classification of the sciences, see Flint (1904).
forces sustaining the discursive formations of a Eurocentric patriarchy” (Wiegand 1998, 190). Hope Olson and others have criticized the DDC hierarchy for its marginalization and exclusion of “groups and topics outside of canonical knowledge” (Olson 1996, 302). Meanwhile, calls to “ditch Dewey” on account of its perceived user-unfriendliness have multiplied in an age of instant keyword searching and browsing.¹⁰

6.0 Evaluation

Quinton (1974, 10) asserts that there are “two kinds of need” which KO systems must serve. “The first is sternly practical. . . . Classification by subject-matter is essential to the reader with access to the [library] shelves to show him what there is on the subject he is interested in.” Quinton continues: “. . . [T]here also exists a theoretical interest in attempting to find some ideal, or at least proper, order for the various fields of knowledge.”

Taking the “theoretical interest” first: There are at least three different aspects of a KO system that we might consider when conducting our test of propriety. These are (a) the extent of the range covered by the entire set of main classes; (b) how and where the boundaries of individual classes are defined; and (c) the ways in which individual classes are related to one another. Simultaneously, there are at least four different kinds of criteria that we might choose to use to test the propriety of a KO system. These are (a) correspondence with some objective reality or ground truth; (b) internal consistency or coherence; (c) utility: i.e., ease, efficiency, and effectiveness of use; (d) morality: i.e., construction in accordance with some code of ethics. In at least the case of utility, then, the theoretical question devolves to the practical one, on the pragmatic assumption that the ideal system is the one that works best.

How should a program of evaluation of the practical utility of these three KO systems proceed? It would be impractical to compare one scheme (or even one version of a scheme) with another, regardless of whether or not the comparison were for the same application context (encyclopedia entries, concepts, library resources). Yet a common evaluation program would be good, if only for the sake of efficiency. Information retrieval (IR) evaluation, based on measurements of query–document relevance, could potentially provide a model;¹¹ but (a) limitations on the kinds of tasks and goals that are involved in IR tests, and (b) a lack of pools of relevance judgments in the contexts in question, mean that other methods should be explored.

One simple plan might be to observe real users engaged in meaningful tasks, and to ask them to rate their success—a test, in other words, of user satisfaction. Given the sense that “going digital” has too often involved throwing the baby out with the bathwater—e.g., OoK’s demise with EB’s move to fully-digital in 2012; SoC’s absence from implementations of online thesauri—an effective design could be to compare “print with KO scheme” vs. “print without” vs. “digital with KO scheme” vs. “digital without.” If such a program were to provide warrant for the reinstatement of “digital with” as a vital part of users’ knowledge-seeking routines, so much the better.

Our fourth criterion, the ethics of scheme construction and usage, has come more and more to the fore in considerations of KO evaluation. Apparently attempting to deflect

¹⁰ See, for example, Chiavaroli (2019).
¹¹ See, for example, Harman (2011).
the kind of criticism that has in recent years increasingly been leveled at the *DDC*, Adler (1974, 6) asks how the OoK can avoid “tendentiousness or arbitrariness.” Does it not “reflect, perhaps even conceal, a commitment to one set of organizing principles rather than another? Does it not embody biases or preconceptions that are not universally acceptable?” Adler provides two immediate responses: (a) that the OoK was constructed “in the light of detailed recommendations, directions, and analytical contributions from scholars and experts in all the fields of knowledge represented”; and (b) that it was conceived, not as a hierarchy, but as “a circle of learning.”

Dorothy Auchter (1999, 295) notes that many of the early reviewers of the 15th edition of the *EB* were “simply bewildered” by its tripartite structure. The decision to retain the alphabetical arrangement of *Macropædia* entries, rather than to follow the topical arrangement adopted in the *Propædia* was widely interpreted as a failure of nerve, one that “seriously undermines” the ability to browse among entries of related interest (Auchter 1999, 295), and that leads to frustration and scepticism. Samuel McCracken (1976, 63) describes the “dismembering” of the *EB* into “mini- and maxi-pedias” as “devoid of benefits” with “nothing to recommend it.”

“The *Propædia*, at least, is harmless,” reckons McCracken (1976, 63). Others are less kind. Suzanne Selinger (1976, 440) is concerned about “the problem of bias or subjectivity,” noting (441) that it is no coincidence that “[t]he *Propædia*, the circle of learning that can theoretically begin anywhere, chooses to begin in its printed appearance with science.” Moreover (442), “it is clear from the weighting of topics that the values of the *Propædia* . . . are [preponderantly scientific].” This bias towards science is ironic given (442) that “[o]bjectivity and neutrality were among the great goals of the scientific method,” and (444) that “objectivity, absolutism, and the unity of truth” are [Adler’s] ideals and beliefs.” In sum (445), the *Propædia* is “grounded in and inseparable from the values, approaches, and presuppositions of scientism.” As a result (445), “The two cultures are not reconciled; one has been opted for at the expense of the other.”

Anthony Quinton (1974, 9) is fairly suspicious of Adler’s credentials: “Adler’s long association with the movement in Chicago . . . which has sought to restore to learning unity of a kind exemplified in the work of Aristotle and with a pronounced neo-Thomist inflection” might reasonably raise questions about potential bias in the structuring of OoK. Nevertheless, Quinton (10) admits, “One would have to be very suspicious to think that this [conception of learning as a circle or “old-fashioned pie”] concealed some deep, ideological design and that perfectly sound and straightforward reasons had not been given for it.”

Educator Robert McClintock (1976), meanwhile, is sharply critical of the 15th edition of the *EB* for its “inadequacy as an educative instrument.” Firstly, a “cult of authority, objectivity, and neutrality” is “embodied” in the *EB*, making it impossible for articles to be included “in which the author concretized and spoke directly to the curiosity and intelligence” of the layperson; secondly, most articles speak in an authoritative voice about fields of established knowledge, rather than in an educative voice about the questions that readers have; and thirdly (and most importantly, in the present context), the OoK, while “impressively complete in its range and detailed in its elaboration,” is arranged in an authoritative rather than a pedagogical order. “In working out an authoritative order, one starts with a body of knowledge and asks what order do
the authorities see in it . . . In working out a pedagogical order, one starts with a student and asks what order he should follow if he is best to apprehend the subject at hand . . .” Whoever takes the OoK at face value and follows it from its beginning, as a guide to self-study, “will be sent first to a long article on the ‘Nucleus, atomic’ . . . written entirely in the authoritative voice . . . simply not a feasible point at which to begin his study.”

Quinton concludes his 1974 review by calling OoK “an immensely thorough and detailed piece of work.” For Quinton (11), “It embodies in its general form no striking innovations and does not conceal within it any principles likely to provoke controversy. Free from architectonic Procrusteanism it seems, for all its elaboration, a practical answer to a practical problem. Only the most exquisitely fastidious could think that they are somehow being got at.” Nevertheless, by the early 1990s, it was uncontroversial for Whiteley (1992, 84) to assert that “There is no index to the Propædia and it is difficult for the unsophisticated reader to use. This volume appears to be the least-used part of the encyclopedia.”

7.0 Conclusion

With the publication of the final print edition of EB in 2010, Adler’s OoK seems to have died a largely un lamented death. Whether or not the OoK, or anything like it, will ever be resurrected and pressed into service online is a matter for Encyclopædia Britannica, Inc.’s accountants to consider. That there remains a market, however shrunken, for authoritative KO systems like the OoK to serve as guides for knowledge-seekers is demonstrated by the publication in 2019 of an 8th print edition of RIT (in the face of a plethora of online thesauri offering instant keyword searching), and biannual print versions of the DDC (in addition to the continuously updated WebDewey service to which thousands of libraries around the world subscribe). Whatever the ultimate fate of the OoK itself, perhaps the simple concept of the “circle of learning” that distinguished the OoK from its predecessors may profitably be salvaged, and used in response to critiques of line- and tree-based systems that necessarily have tops and bottoms, firsts and lasts, beginnings and ends.

One additional function of KO schemes that has so far gone unremarked, and that goes beyond even the pedagogical function mentioned above, is one that we might call the generative function. What does one get out of reading the OoK, the SoC, or the DDC? A sense of one particular view—perhaps fundamentally mistaken—of the shape, size, and structure of the totality of knowledge, for sure; but also, potentially, ideas about new fields, new concepts, new subjects that are not currently part of that totality, and ideas about new ways of organizing that totality. The generative function is what continues to make the future of KO system design so exciting.

References


Advancing the Interoperability of the GLAM+ and Cultural Tourism Sectors through KOS
Perspectives and Challenges

Abstract:
The possibilities and challenges of knowledge organization systems (KOS) to collaborate in the interconnection between the cultural heritage sector–galleries, libraries, archives, museums, publishers… (GLAM+–) and the increasingly important cultural tourism industry are explored, and a model for framing their interaction is proposed. Due to the diversity of KOS implied in GLAM+, this project is to be treated as an interoperability problem, thought a strong user-oriented purpose is also needed, based on a careful assessment of tourists’ segmentation and their needs. The main components of the model are five: the real phenomena that form up the potential world of interest, the universe of potential sources, a web taxonomy, a domain thesaurus and an interoperability hub. The St James’ Way is used as a source of examples. It is concluded that thesauri based on ISO 25964 offer a great potential for the simple, flexible, dynamic and distributed interconnection between the institutions of memory (GLAM+, research institutions in digital humanities and social sciences, transparency portals…) and the growing demand from the tourist sector for a more personalized and contextualized experience that can make a difference in an increasingly competitive international market.

1.0 Introduction: context and motivation
It has become a common experience that web information is changing tourism and travelling, but a new phase is developing in the last years. There is increasing evidence of “a growing “bifurcation” between traditional online travellers, i.e., those who use the Internet for standard travel products, and those who are beginning to adopt alternative channels and products in search of deeper and more authentic experiences”, with the first market entering into a relative stagnation and the second offering new opportunities for combining different products (Xiang et al. 2015). Some experts have even identified a ‘cultural turn’ in tourism (Dabbage 2018a, 55-56; 2018b).

In this context, relating the immaterial, material, artistic, bibliographical and archival heritage to competent touristic proposals and infrastructure to be used by people of different nationalities, languages and cultures is becoming a key strategical challenge for both sectors. On one hand, the tourism sector can profit from more contextualized, personalized and interesting information resources. On the other, Humanities and GLAM+ (acronym for galleries, libraries, archives and museums, and, in general, the heritage preserving institutions) can improve their visibility and relevance through an undisputable, applied and practical contribution to social and economic development.

Rich references to cultural artefacts–both before, during and after the visit (in webpages, VR applications, QR code support, augmented reality…)–can help tourists to opt for a particular tour; decide future activities outside the standard ones, widening choices and benefiting the local tourism market; improve their cultural, educational and life experience; promote word-of-mouth recommendation; and enhance the acquisition
of cultural and historical lessons and knowledge. Such a vision seems a gain-gain one both for tourists, destinations, and tourism agents and organizations.

Within this framework, this paper explores how experts from the field of knowledge organization and information architecture can collaborate on devising and proposing action lines to improve the feedback between both sectors (tourism and cultural heritage institutions). In a first stage, by facilitating the use of the huge amount of data and digital artefacts made available by the information and communication professionals, humanists and social scientists, which allow the contextualization, enrichment and personalization of the touristic experience in its relation to relevant cultural objects. Second and reciprocally, by looking for strategies to enhance the transfer of resources from the tourism industry to the field of basic research in the humanities and social sciences, promoting its economic sustainability by connecting it with its potential market uses. This last point seems specially important at a time when Humanities have become neglected by funding agencies because of the longstanding economic constrains following the 2008 crisis and the increasingly acrimonious cultural wars between global and identitarian political stakeholders.

2.0 Aims and research questions

The overall intention of this paper is to explore how KO research and development can contribute to the interconnection of the institutions of memory (libraries, archives, museums, documentation centres, research institutions in digital humanities and social sciences, transparency portals…) and the needs of cultural tourists, an increasingly important industry (Fang 2020); and to develop a model that can contribute to frame and guide future research on the field. Specifically, six research questions were addressed: Which roles can the GLAM sector perform inside the digital information ecology of cultural tourism? Which are their implications for KO? How can the relation between the GLAM and cultural tourism sectors be modelled to reveal the central role of KOS in such interactions? What are the relevant characteristics of KOS used in the main GLAM subsectors? Are they interoperable with cultural tourism websites? On which terms? To answer these questions, in the following sections, the main agents and factors involved in the interoperability between the GLAM and cultural tourism sectors are identified and considered from the point of view of Knowledge Organization, that is, of the role that KOS might have in their successful interaction; and a model is proposed. Finally, the problems and opportunities for KO in this field are identified.

4.0 Analysis and discussion

In the next sections, we will consider the relation among cultural tourism and the GLAM+ sector in the digital age as a potential information ecology. In such an ecology, three main kind of agents and artefacts can be identified: tourism information mediators and their web sites; cultural heritage institutions (GLAM+ sector) with their information sources and databases; and end users (tourists and travellers) with their needs, which are partially known but must be partially disclosed.

In this emergent ecology, KO can contribute to close the gaps among their agents in three ways: modelling the connection among GLAM+ resources and user needs; identifying knowledge representation and organization technologies and methodologies that
can be useful to implement the model; and offering an operative model to start experimentation. Two contextual problems outside the model must be also stressed because of their current importance: communicational issues and legal concerns.

4.1 The information ecology of cultural tourism

The concept of ‘information ecology’ allows the modelling of information systems that have not been purposely designed, in contrast to those compact and well-differentiated from their environment, like libraries, archives and information centres. An information ecology (Hubermann 2001; Shim and Lee 2006; Sebastiá 2008) can be defined as a network of information organisms interacting among them and with their environment to form a complex system. This concept is very useful to think about situations where different information agents co-exist, cooperate and compete to fulfil an information need. This is the typical situation in the Internet; and, in our case, what users interested in cultural tourism will experience when trying to solve their information needs.

For example, Internet information on the St James’ Way is provided by a distributed network of independent agents: the non-governmental sector (associations, religious institutions, informal groups and individuals), private for-profit companies (travel services firms, publishing houses, consumer cooperatives...), and public institutions (council, national and regional governments...) (López del Ramo and García Marco 2018). Each of them has its own aims and provides specific information services to pilgrims and tourists, who, on their part, are also very segmented regarding their specific interests.

4.2 The tourist information mediators (providers and associations)

Though GLAM+ institutions sometimes address proactively the information needs of tourists, relevant information is generally vehiculated through the tourism industry, and more precisely by those institutions and departments that are specialized in connecting tourists and destinations. This is mainly a marketing activity, and their more specialized agents are the so-called destination marketing organizations (DMOs): institutional, council, regional, national tourist information offices, and their background marketing departments. DMOs can be the main mediator agents between GLAM+ institutions and tourists, though online travel agencies (OTAs) and social media have been gaining prominence (Xiang et al. 2015). Of course, it should not be forgotten that there are big GALM+ institutions that have great DMOs inside them, e.g., big museums and galleries. Associations of experts and fans related to the protection and dissemination of cultural heritage are also another important sector of mediators in the field of cultural tourism. Internet has brought a revolution to marketing, and now even the smallest organizations can have an Internet global presence.

From the point of view of tourist mediators, knowledge organization may have two different uses, one internal and the other external. The internal objective is related to knowledge management: the representation of knowledge into information; information preservation, retrieval and sharing; and the transformation of information into knowledge. For this, a corporative KOS is needed.
The external goal is to communicate the part of this information that is relevant to tourists, so that they become aware, transform it into knowledge and hopefully choose the proposed destinations for their travels. This is usually done through websites and social platforms, but also increasingly by more and more sophisticated mobile applications. In this regard, tourist information systems must be seen as integrated, with a core of organized information and a set of well-established distribution procedures. The main KO tools for these purposes are web taxonomies in the case of cultural tourism webs, and folksonomies and ad hoc taxonomies in the case of social networks and blogs.

Of course, knowledge organization experts now that, for proper functioning, both aims must be connected into a successful knowledge organization system, though both of them should be efficaciously addressed and easily differentiated. In fact, both functions are more or less separated in many organizations: sometimes one is absorbed or displaced by the other; other times website and corporate information databases are not well communicated. Frequently, the corporate taxonomy becomes the structure of a tourist website, and, as a result, the website reflects more the ontology of the organization than that of tourists (López del Ramo and García Marco 2018) or is mainly product-oriented; but there are also many well-designed sites from a user-oriented perspective (Table I).

<table>
<thead>
<tr>
<th>Table I. Main categories in the taxonomies of four St James’ Way websites</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>caminodesantiago.gal</strong></td>
</tr>
<tr>
<td><strong>Galicia government tourist-oriented website</strong></td>
</tr>
<tr>
<td>Discover</td>
</tr>
<tr>
<td>Get ready</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Plan your trip</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>On the way</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Knowl. and research</td>
</tr>
<tr>
<td>Xunta de Galicia</td>
</tr>
</tbody>
</table>

4.3 The side of the information sources: GLAM, research centres and publishing

The information sources side is quite complex. Any approach to integrate GLAM resources in tourist information sites requires enhancing interoperability among very diverse systems.

Libraries are by far the more standardized GLAM subsector, though the gap is being closed very quickly because of the pressure for global and integrated access that Internet at the same time offers and requires. Their resource description and interchange standards are fully international and integrated (MARC21 family), and to a lesser grade also are their KOS, both systematic classifications (LCC, DC, UDC…) and alphabetic subject headings lists (LCSH, RAMEAU, EMBNE…). There has been even a strong work on integrating systematic and alphabetic KOS in libraries, some finished and other still
on course (LCC-LCSH, UDC and national subject heading systems in Europe…). Also, relevant mappings among competing systems are on course (e.g., Slavic 2011). These are strong points for the interoperability of the GLAM and cultural tourism sectors. But there are also weaknesses from the tourist or traveller perspective. Classifications frequently do not provide the level of specificity required to map sources to their needs; and topics are scattered among many classes, doing mapping projects really complicate.

For example, there is no class in the UDC for St James’ Way. The classification of works about the St James’ Way is usually done in the Regional Geography class (913), and less in Routes, etc. (656.022) or in the many (and scattered) classes available for Travel (e.g., 338.48, 656.022.33), as transport and tourism are in different trees. For the purpose of adding specificity, this selected class is frequently faceted in Spain by the two main countries: (44+460), that is, France and Spain (to denote the “Camino Francés”), though the actual countries may change (e.g. Portugal). The National Library of Spain has noted the lack of specificity of the expression, and its librarians have usually added “Camino de Santiago” between the brackets and after the country codes.

Subject headings are usually much more specific. Though the exact subject heading is not present in many international systems, it can be easily recognized in the strings:

- EMBNE: Peregrinaciones cristianas — Santiago de Compostela
- LCSH: (Christian pilgrims and pilgrimages-Spain-Santiago de Compostela)
- RAMEAU: [Pèlerinages chrétiens-Espagne-Saint-Jacques-de-Compostelle (Espagne)]

Therefore, there is a strong potential for an easy interconnection of libraries and cultural tourism websites. But a serious problem stands in the way: practical travel decisions require data, not systematized knowledge, and there is no easy method to transform books and other complex library materials into high quality data without the cooperation of the research, publishing and media sectors, as it will be discussed later.

Big museums, which rely on very analytical databases for the control of their collections and operations, are at the forefront of providing semantic data that can be linked by external sites. This data provides factual, contextual and bibliographical information on pieces of art and craftsmanship and reproductions that can be inserted in websites and A/VR applications. The CIDOC CRM ontology, originated in the museum field, provides a full frame for the interoperability of other controlled vocabularies and is being influential in the GLAM sector. Museums have also sound KOS: AAT, Iconclass, etc.

Archives, their fonds and documents are rarely used by tourists and their mediators in a direct way. There are interesting projects linking historical documents to heritage sites, monuments and pieces of art, but this kind of information is usually secondary for tourists: they will use it only out of curiosity and not for practical decisions. Archives are also very idiosyncratic in their intellectual organization, as their KO leitmotiv is the principle of provenance and respect for the original order, which is complemented by the macrolevel organizations based on function to allow the navigation of the multiple changes that organizations suffer along their life. Anyway, authorities have grown increasingly important for the multilevel description in archives, and there is a strong movement to make library, museum and archival headings compatible.

Outside GLAM but closely related, there is a huge industry that is also fully involved in knowledge preservation: the publishing sector. Though a first sight could suggest that
their role is incorporated through the library network, editorial organizations have a huge task ahead that is crucial for the successful interoperability among the cultural and scientific sectors: adapting the publishing practices to the promises and requisites of the semantic web and big data revolutions. Cultural tourists have two broad categories of information needs: a) gaining broad perspectives for travel planning and background knowledge development, and for this purpose books and journals are totally suitable; b) acquiring and accessing data throughout different sources, and it is here were the digitally-enhanced user has problems. Most books are oriented to the first need; and, as they are not usually automatically searchable and navigable through library systems, a big gap is opened between both worlds. Semantic web offers the technology to bridge this gap, but editions and library catalogues must be transformed to fulfil this promise.

What it has been said about the ‘traditional’ publishing sector must be applied to the current multimedia environment, including audio-visual industries and videogaming. Other emerging ones, like augmented and virtual reality, use semantic technologies by default.

Strictly, the GLAM acronym does not include the producers of information; and, as a great part of the task of producing semantic- and data-supporting documents rely on them, it is suggested that an extended superset should be considered to include publishers and other media producers, e.g. GLAM+.

4.4 The tourist side: end users first?

It has been observed that “huge discrepancies exist between the domain ontology derived from tourism Web sites and the one emerging from user queries”, mainly because the tourism industry uses a very specific terminology and categorization that is not simply connected with the terms that users actually search, and because, apart from a group of overrepresented categories, the “the overall domain is extremely rich and largely idiosyncratic […] with numerous destination specifics” (Xiang et al. 2009). Though consumer generated contents – e.g., blogs and reviews (Gretzel, Hwang, and Fesenmaier 2006) – can be used to learn about the language that travellers use, filling the terminological gap; there are problems that do not seem easy to solve, especially those derived from the intrinsic limitations of attention, processing memory, interface space, and the representation of a really complex domain.

Problems increase when taking into account that cultural tourism is a huge sector and that tourists themselves are indeed very segmented. Only in the field of cultural tourism, McKercher and Du Cross (2002) identified five types of tourists: purposeful, sightseeing, serendipitous, casual and incidental; and Csapó (2012) classified seven different kinds of cultural tourism products for them: heritage tourism; cultural thematic routes; cultural city tourism, cultural tours; traditions, ethnic tourism; event and festival tourism; religious tourism, pilgrimage routes; and creative culture, creative tourism.

On the other hand, this gap between the overwhelming worlds of information and tourists is a very interesting challenge that has attracted a lot of talent. In particular, modelling users through ‘ontological’ profiles, based on types and/or personality traits, is an effort that has been on course for more than a decade and now constitutes the established research topic of tourism recommendation systems (Grün, Neidhardt, and Werthner 2017). However, GLAM+ connection requires specific typologies for both
tourists and sources, based on their information needs: broad ones—planning, and connecting and developing their mind maps—and specific ones—solving problems with relevant data, and navigating among sources in search of unanswered questions—.

In this situation, mediators and users can work through general-purpose search systems like Google, or with the source providers’ KOS. In the first case, users will obtain relevant selections, but not necessarily very precise when leaving the first results, neither exhaustive or filtered by source. With the second approach, these aims could be better served, but a difficult work of knowledge organization engineering will lie ahead. The first evidence gained in any experience in KOS/GLAM+ interoperability for cultural tourism is that it is no easy task, specially when leaving the basic operations and data and getting into the realm of learning, sense-making and culture.

4.5 Bridging the gap (1): towards an ontological and epistemological model

Thus, can a model be developed to integrate the different information needs, structures and finally perspectives of GLAM+, DMOs and tourists so that a truly cultural tourism information ecosystem may be born?

In our opinion, any proposal has to take in account four different layers: the cultural world (artefacts, persons and organizations, sites, abstract realities), counting on the KOS that are already functioning in the GLAM sector; cultural tourism ‘science’ and operation, as expressed in some KOS, like the important WTO thesaurus (World Tourism Organization 2001), in the tourism website taxonomies and some prospective ontologies that are being developed (Li, Buhalis, and Zhang 2013); and the last one for representing the user. To this last aim, Maslow’s (1954) pyramid of human needs seems especially well suited to frame the classification and interrelation of so different layers as basic needs (accommodation, eating and drinking, security…), social ones (identity, relation…) and self-actualization (sense-making, high culture, wisdom-building…).

4.6 Bridging the gap (2): available standards and technologies

To model such a KO ecology, fully developed technologies and standards are available both in the field of KOS and in the semantic web realm. Regarding networking knowledge organization systems (KOS), thesauri offer a great potential for the simple, flexible, dynamic and distributed interconnection between the institutions of memory, especially after their relaunch with the new ISO 25964 standard (Aitchison and Dextre Clarke 2004; Dextre Clarke 2012). In particular, Part 2 of ISO 25964 deals with the interoperability of all these different KOS implied in the information ecology of cultural tourism, and proposes specific mapping models and specific devices (=EQ, ~EQ, +, |, Bm, NM, RM). In our project, we are experimenting with a hub architecture, but a lot of system-to-system projects are on course, so the environment is changing quickly.

Dextre Clarke (2011) has made a very clear diagnostic of what are the expected results of mapping thesauri and the main kinds of KOS, and Soergel (2010) has proposed a clear conceptual framework that also considers facet-based search. In the particular case of cultural tourism, more positive results can be expected from interoperability with authority and subject lists, because they accurately represent cultural artefacts (through titles), persons and places; though the problems will persist for higher level concepts.
On the part of the semantic web research, W3C has by now greatly completed the standards that deal with its lower and middle layers; thesauri have been successfully expressed in RDF, OWL and SKOS; and tourism recommender systems based on semantic web and ontologies have become one the main research fronts in information science and tourism: “Recommender systems based on semantic web and ontology technologies are an effective method and tool to improve the quality of internet service through personalization and customization.” (Li, Buhalis, and Zhang 2013).

4.7 Bridging the gap (3): a methodology

KO interoperability is difficult when working with already well-seasoned KOS, but it becomes a mess when trying to connect ad hoc web taxonomies and folksonomies—on the side of DMOs and other traveller-oriented websites—with GLAM+ KOS on the other. In a current project on the Aragonian part of the St James’ Way, we have divided our work in two lines of action.

On one hand, the different websites are being classified in homogeneous groups and their taxonomies studied: council, regional and national DMOs, mainly from the public sector; pilgrim associations; and business providing information (publishing houses, consumer cooperatives, hotel chains, transport firms…). In this way, it is possible to obtain some sort of a ‘least common multiple’ of all these empirical taxonomies, including all their concepts without repeating them. Equivalent terms, which are extremely important in the “language of tourism” (Xiang, Gretzel, and Fesenmaier 2009), are then controlled, so they can be used to resolve searches; taxonomy-incompatible polyhierarchical relations are resolved into BT/NT and RT; and the relations among the taxonomy categories (now thesaurus concepts) are expressed through RT. As a result, a special kind of thesaurus emerges, which could be called a Common Compact Taxonomical Thesaurus (CCTT). Each of its concepts is to be a node (page) in a trial CMS-supported website. Thinking in future interoperability, a SPARQL point should be developed, with an independent system behind to manage prospective and external KOS links.

On the other hand, the most used KOS and knowledge representation systems in the different GLAM+ sectors are being studied to explore their potential interoperability. Prospective analysis has shown that mapping each KOS to the others would be a difficult, uncertain and huge task; and that it is also unfeasible to take one of them as a hub. Therefore, the idea is to connect the subsets of them that offer greater potentiality (because they are related to the concepts of the CCTT) to the nodes of the website.

Though the best solution for actual interoperability would be to establish an independent KOS hub, for the moment we intent to build a trial DMO website that will incorporate the pilot CCTT as an extended taxonomy; and then link the external resources to each node of the CMS-based website, based on a case-by-case analysis of their correspondences with the controlled tags in the sources. Complementarily, relevant Wikidata nodes and their relations could be mined to produce a set of mappings.

All these steps are based on a presupposition: that DMOs KOS really express user needs, but this extreme should be further researched, because actual evidence is limited. This gap can only be filled in the future by doing user studies: both subjective (satisfaction surveys, etc.) and objective (eye tracking, log analysis, search analysis…). Certainly, the strong need to connect both subdisciplines of information science (i.e., KO...
and user studies) becomes even greater when approaching multidisciplinary and multi-platform fields, like cultural tourism and GLAM+ integration. Only in this way, the whole cycle of evidence-based KOS design and research could be effectively closed.

4.8 Beyond KOS and linked open data: legal concerns and communication issues

Although the paper is focused on the analysis of networking knowledge organization systems, there are other aspects that are essential for their operations in the real world: in particular, communicational issues (communication security, effectiveness of the informational message) and legal concerns (privacy, data protection and intellectual property, basically), especially considering the problems created by the massive data processing that is becoming inherent to the Internet, which requires the use of algorithms that respect the legislation and standards on human rights in the use of open data.

5.0 Conclusion

In this paper, it has been shown that KO, cultural tourism, GLAM+ and humanities can develop a fruitful alliance as disciplines. In the digital realm, cultural tourism needs humanities to support sound and personalized products, incorporating GLAM+ resources and data to enhance tourists’ experiences. To integrate all this information, it has a networking KO gap. On the other hand, cultural tourism is a fascinating subject for KO. First, it is a truly transdisciplinary field, while far from compact. Second, it brings forward the problem of end users’ segmentation with their different information needs, and therefore diverse conceptual maps.

In practice, a KO-enhanced operative model for the networking of GLAM and cultural tourism websites has been outlined, generalizing from an on-going project on the St. James’ Way. As a conclusion, it can be affirmed that networking KOS—in particular thesauri following its relaunch after the new ISO 25964 standard (Aitchison and Dextre Clarke 2004; Dextre Clarke 2012)—offer a great potential for a simple, flexible, dynamic and distributed interconnection between the institutions of memory (libraries, archives, museums, documentation centres, research institutions in digital humanities and social sciences, transparency portals…) and the growing demand of the tourism sector for a more and more personalized and contextualized experience that can make a difference in an increasingly competitive market (Abrahams and Dai 2005).

Acknowledgments

This research work was carried within the CSO2015-65448-R (MINECO/FEDER) project. We are also grateful to the anonymous referees for their suggestions.

References


Ann M. Graf – Simmons University, United States

Domain Analysis of Graffiti Art Documentation
A Methodological Approach

Abstract:
Details are presented of a recent research project undertaken to ascertain the documentary and descriptive practices associated with graffiti artwork from within the graffiti art community as evidenced by 241 graffiti websites. Domain analytic methodologies following a pragmatic approach to knowledge organization and using evidence obtained from within an artistic community are extremely useful ways to provide insight into what are the most important facets of information to capture for works not often documented from within libraries, archives, and museums. This paper will discuss various methods used to analyze community-driven graffiti art collection, organization, and description in the online environment, the results of which form a part of the basis upon which a faceted KOS can be built.

1.0 Introduction
Knowledge organization systems (KOS) in use for the documentation and description of artworks have a respectable, if shorter, history compared to those used in libraries (Urban 2014). This is often understood to be due to the fact that museums are most often collecting and creating surrogate records for unique objects (Taylor 1999). Unlike the library, the art museum is representing objects that would not benefit from shared cataloging practice, though this is changing as images of artworks are increasingly available online and users desire to have access to them regardless of where they are physically located.

For certain types of art that often fall outside the purview of the formal institution, such as graffiti art, documentation and organization of the resulting image records is carried out largely by the graffiti art community itself, including graffiti artists and enthusiasts. This is due to the extra-institutional nature of the artworks, the legal complications often surrounding their creation, and the inability to monetize the works when found “on the street” (Schacter 2014). Interest in the artworks continues to rise, despite these documentary challenges, evidenced by the large number of websites dedicated to preserving the images of the works around the world. While KOSs in popular use in the library, archives, and museum environment do not include granular terminology to address the many facets of graffiti art, at least one, the Getty Art and Architecture Thesaurus (AAT), has recently added a limited number of graffiti art descriptive terms.

This paper introduces research conducted on the state of descriptive and organizational practice applied to collections of graffiti art images in websites from around the world. The data for analysis comes from a set of 241 graffiti websites. This stage of the research will report on the categories – or facets – used to organize the image galleries themselves. This research does not explore description as applied to individual images. Knowledge of organizational practice among those documenting the art form in image galleries online is foundational to understanding the implicit categories, vocabulary, and details associated with a particular artistic tradition and may inform further research on documentation and organization of outsider art.

It is very hard to pin down a community, to draw lines around those within and those without. In the case of graffiti art, with its intersecting boundaries of what is legal and what is not, and what is defined by some as true graffiti, and by others as vandalism,
and yet others as the more sanitized term street art, it can be impossible to find consensus. Should the artists themselves be the ones to say what they are doing, how they are doing it, and how it is described, documented, and organized? Are those who simply love the often bright, complex, and (sometimes) publicly placed works allowed to have a say? Do those who actively look for the works and photograph them, sharing them online in large galleries, belong to this community? Because graffiti image galleries online are very often cooperative endeavors, whether knowingly or not, willingly or not, of images submitted by artists who created the works, photographers who stumbled upon them, webmasters or social media mavens who enjoy them, and any combination of these and more, the easiest way to begin this research was to use the evidence of the collections themselves.

It is acknowledged from the start that this is a messy endeavor, trying to decide who is acting and what role they may play in the organization of graffiti art online. While few authors have addressed the challenges of documenting graffiti art specifically (see Masilamani 2008 and Gottlieb 2008 for two of the more robust examples), the process is actively taking place in a very broadly distributed fashion, each participant seemingly acting independently of the others. Despite their autonomy, those doing the documentation in this research represent a community or a domain in that they share “an ontological base that reveals an underlying teleology, a set of common hypotheses, epistemological consensus of methodological approaches, and social semantics” (Smiraglia 2012, 114). This work fills a gap in the research by illuminating the facets for organization of graffiti art images as used by those working to share large collections of the works online.

The value of domain analysis as a research tool is well documented in the knowledge organization literature (Hjørland and Albrechtsen 1995, Hjørland 2002, Smiraglia 2015, Albrechtsen 2015), as well as facet analysis (Hjørland 2013, Cho et al. 2018, Campbell 2004). There are no domain analyses that examine modern graffiti art image documentation and the facets used as attributes to organize the images. Research reported by Graf (2016) revealed the lack of graffiti art-related terminology available in the AAT. Only three out of the twenty most often used graffiti terms from her analysis of graffiti zines appeared in the AAT. Interestingly, within two years, eleven more of the same twenty terms were added to the AAT, bringing the original percentage from 15% to 70%. This indicates the influence of the graffiti art community and their practices, and the reporting of research on those practices, on widely used professional tools for the documentation and organization of artworks. Further granularity can be found in this current research for those desiring to extend the available terminological offerings for graffiti art documentation.

2.0 Methodology

As preliminary research in this area, the first step was to decide exactly what would be examined. There are many websites devoted to the documentation of graffiti and street art. One of the very first and most well known of these is Art Crimes (www.graffiti.org). The About page on their website states that “Art Crimes was the first graffiti site on the net, and we're still one of the biggest …” (Art Crimes 2020). As an early and large graffiti art website, Art Crimes has gathered links to numerous other graffiti and street art websites around the world. At the time of the research, Art Crimes included a
list of 709 links to other sites. This list was used as the basis for the eventual set of 241 websites evaluated.

Each of the 709 links on Art Crimes was visited and a judgement was made on whether or not to include the site in the study based on several criteria. 318 of the links were either dead, empty, or presented a notification that the site had moved without providing forwarding information. 64 sites were fully in languages other than English and therefore eliminated from the study, though some sites that were kept employed other languages but kept navigation labels for the site in English. 57 of the sites were professional artists’ sites, not specifically galleries of graffiti or street art images. 20 of the sites were not relevant because they were focused on music, advertising, or other products or services. Eight of the sites were links to social media galleries, such as Flickr or Instagram. These were not included in this study at this time because of the organizational confines of social media platforms. Each social media platform includes specific ways that uploaded images can be labeled, organized, and grouped. There exist thousands of graffiti and street art image galleries on social media platforms, and they are ripe for further investigation, but were considered outside the purview of this research. One site among the 709 was not an independent website, but rather a sub-page of the Art Crimes website itself, and therefore eliminated, though Art Crimes itself remained in the study. After all of the sites were evaluated in this way, 241 live sites remained.

Each site was evaluated for structural elements of pages and sub-pages, indicated by navigation labels and hotlinked text. Examples of navigation labels can be easily seen across the top of the webpage banner in Figure 1 for the site 50mm Los Angeles (http://www.50mmlosangeles.com/). These navigation labels include: Gallery, Articles, Events, L.A. Legends, Blackbook, Links, Forum, About Us, and Submit an Event. There is also hotlinked text, “login | register”, that leads to other sub-pages of the website. All of these labels were entered into a QDA Miner database for each of the 241 websites.

![Figure 1. 50mm Los Angeles website home page with navigation labels.](image)

Each of the sub-pages accessed through the navigation labels and hotlinked text were visited and each was evaluated for evidence of further navigation labels, or sub-divisions of organization. Some websites had very shallow organization with only a couple levels, while others were deeper structures with several levels. An example of a website with deeper structure is Fatcap (fatcap.com). The home page of this website includes a navigation label for “pictures.” Hovering the mouse over this label gives the user several sub-levels from which to choose: all pictures, worldwide graffiti, artists, crews, types, supports, and styles. Clicking on “worldwide graffiti” takes the user to a new sub-page.
that includes links to 117 deeper sub-pages, arranged by larger geographic regions: Africa, Asia, North America, South America, Europe, and Oceania. Clicking further, on “United States” for example, takes the user to a list of 40 states and the District of Columbia, each with from 1 to 69 individual cities linked as sub-pages to go even deeper.

Other websites only had a couple levels, but then divided a single level into hundreds of sub-pages. An example of a sub-page in a shallower organizational structure with numerous organizational divisions, again from the website 50mm Los Angeles, is shown in Figure 2. Only part of the screen is visible in this image, but the organization of the image gallery is divided into an alphabetical, hotlinked list of navigation labels, some representing artist names (pseudonyms), locations, and styles, etc. Clicking on any of the hotlinked labels in this expansive list will take the user to a gallery of graffiti and street art images with works that have something to do with the label. Each of these sub-pages used as organization for image galleries was also entered into the QDA Miner database for further evaluation, as will be explained in detail below.

Once all 241 websites were visited and all navigation labels were entered into QDA Miner for all levels of organization, each individual label was coded to indicate what type of organization was indicated. The coding developed as the analysis proceeded. Six broad categories of codes, or facets, evolved during the research, two of which focused
on the websites themselves, and four of which focus on the artwork images on the websites.

The two categories of codes that apply to the websites themselves include Sites and Other Media. The Sites category includes navigation labels that refer to aspects of the websites, how users can interact with the websites, and other information related to shopping, subscribing, and other graffiti and street art-related information accessed within the websites. They do not concern description of graffiti and street art images associated with the image galleries on the websites. The Sites codes include: About, Contact, ContributeFlix, Disclaimer, FAQ, Forum, Glossary, Guestbook, History, HowTo, Interviews, Map, MyAccount, Poll, Shop, Subscribe, and Videos. This category of codes relates to the structure, navigation, and use of the website in general. The second of the two categories not concerned with description of graffiti and street art images is the Other Media codes. This category of code was applied to navigation labels that linked to a blog or social media account associated with a website, such as an Instagram, Facebook, or Flickr account, or to a list of links to other graffiti or street art sites, or other associated media located outside the websites studied.

The remaining four categories of codes are the focus of the research reported herein. These categories were used to describe graffiti and street art images themselves and include General, Types, Supports, and Locations. Each of these four works-based categories will be described in greater detail and will provide insight into the documentation, description, and organization practices of the graffiti art community, which includes artists, photographers, and various enthusiasts as described earlier.

3.0 Findings

Each of the code categories is divided into subcategories, which reflect aspects of description for graffiti works. The first of these is the General category, which is divided into 17 codes, or facets, as shown in Table 1. In each of the code tables, the name of the code is given first, followed by how many times that code was applied over all 241 websites. The third column indicates the percentage of all codes applied. The fourth column indicates how many of the 241 sites earned that code at least once, followed in the last column by the percentage of all sites that used that code at least once. Some of the websites earned the same coding in multiple places on the site, which accounts for the sometimes very large number of individual codes, like Artist. Whenever the name of an artist was used as a way to organize a gallery of images, that label of the artist’s name was coded as Artist. As evident in Figure 2, some websites included hundreds of individual artist’s names.

Each of the tables lists codes in order of the percentage of sites making use of the code at least once, providing a type of ranking for the popularity of an aspect of organization across all sites. This also avoids the skewing effect of using each code instance as a popularity measure instead. While the month of a work was applied 35 individual times, putting it fifth in terms of instances, it was seen on only 2.1 percent of all sites, or 5 sites, which indicates it was 13th out of 17 in popularity. Some of the General codes reflect common aspects of traditional art documentation, such as the use of an artist’s name or the year. Others reflect affordances of an online gallery, such as New, Color, Featured, RatedHigh, and Old. A very interesting aspect of many of the General codes is their specific applicability to graffiti art. This is evident in
codes such as Gallery (used to indicate when a work was in a gallery and not in a traditional graffiti location), RIP (used for commemorative pieces in honor of a graffiti artist who has died), Legal, Outside, and Illegal. Most graffiti is assumed to be illegal, but there are also legal walls where graffiti is allowed. It makes sense to include organization for legal works, as they are created under very different circumstances than illegal ones. It doesn’t make as much sense to offer organization specifically for illegal works, and this code was applied on only 2 sites, compared with 7 sites that earned the Legal code.

Table 1. General codes and their usage across all sites.

<table>
<thead>
<tr>
<th>General Codes</th>
<th>Count</th>
<th>% of Codes</th>
<th># of Sites</th>
<th>% of Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artist</td>
<td>14439</td>
<td>71.2</td>
<td>50</td>
<td>20.7</td>
</tr>
<tr>
<td>Event</td>
<td>89</td>
<td>0.4</td>
<td>31</td>
<td>12.9</td>
</tr>
<tr>
<td>Gallery</td>
<td>49</td>
<td>0.2</td>
<td>29</td>
<td>12.0</td>
</tr>
<tr>
<td>Year</td>
<td>227</td>
<td>1.1</td>
<td>27</td>
<td>11.2</td>
</tr>
<tr>
<td>New</td>
<td>35</td>
<td>0.2</td>
<td>26</td>
<td>10.8</td>
</tr>
<tr>
<td>Old</td>
<td>35</td>
<td>0.2</td>
<td>26</td>
<td>10.8</td>
</tr>
<tr>
<td>Featured</td>
<td>27</td>
<td>0.1</td>
<td>20</td>
<td>8.3</td>
</tr>
<tr>
<td>Inside</td>
<td>11</td>
<td>0.1</td>
<td>10</td>
<td>4.1</td>
</tr>
<tr>
<td>RIP</td>
<td>75</td>
<td>0.4</td>
<td>10</td>
<td>4.1</td>
</tr>
<tr>
<td>RatedHigh</td>
<td>14</td>
<td>0.1</td>
<td>8</td>
<td>3.3</td>
</tr>
<tr>
<td>Legal</td>
<td>15</td>
<td>0.1</td>
<td>7</td>
<td>2.9</td>
</tr>
<tr>
<td>Outside</td>
<td>7</td>
<td>0.0</td>
<td>7</td>
<td>2.9</td>
</tr>
<tr>
<td>Month</td>
<td>35</td>
<td>0.2</td>
<td>5</td>
<td>2.1</td>
</tr>
<tr>
<td>Color</td>
<td>12</td>
<td>0.1</td>
<td>4</td>
<td>1.7</td>
</tr>
<tr>
<td>Day</td>
<td>5</td>
<td>0.0</td>
<td>4</td>
<td>1.7</td>
</tr>
<tr>
<td>Decade</td>
<td>8</td>
<td>0.0</td>
<td>4</td>
<td>1.7</td>
</tr>
<tr>
<td>Illegal</td>
<td>5</td>
<td>0.0</td>
<td>2</td>
<td>0.8</td>
</tr>
</tbody>
</table>

The next category of work-related codes are the Support codes. These codes were applied when organizing by the surface upon which the artwork was created or placed. One distinction in this group of codes is the Canvas code, applied here to works produced in a studio. One-third of all sites earned this code, reflecting the difference in perception of graffiti-style artworks committed on canvas as opposed to walls, trains, or other publicly accessible spaces. The use of the street is important to the notion of graffiti and street art (Austin 2010, Riggle 2010). Painting on canvas is often seen as a desire for profit, a safe way to make art in the comfort of a studio, or a type of selling out of the art form (Jacobson 2017). This conception is common enough that a relatively large number of sites used this type of organizational label to separate out works made in a studio from those made on the streets.

Table 2. Support codes and their usage across all sites.

<table>
<thead>
<tr>
<th>Support Codes</th>
<th>Count</th>
<th>% of Codes</th>
<th># of Sites</th>
<th>% of Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canvas</td>
<td>109</td>
<td>0.6</td>
<td>77</td>
<td>32</td>
</tr>
<tr>
<td>Walls</td>
<td>107</td>
<td>0.5</td>
<td>65</td>
<td>27</td>
</tr>
<tr>
<td>Trains</td>
<td>253</td>
<td>1.2</td>
<td>51</td>
<td>21.2</td>
</tr>
</tbody>
</table>
One of the most interesting categories of codes is that devoted to types of art. This category is rich with terminology, much of it specific to graffiti art. It is also the largest of the code categories, with 31 individual codes. Many of the codes are familiar terms that could be associated with more traditional art forms, such as Sketches, Murals, Stencils, Posters, and Political. Many others have specific meaning within the graffiti art community, such as Tags, Pieces, Bombs, Throwups, Productions, TrainWholecars, TrainEtoEs (end-to-ends), TrainTtoBs (top-to-bottoms), Wheatpaste, and Wildstyle.

### Table 3. Type codes and their usage across all sites.

<table>
<thead>
<tr>
<th>Type Codes</th>
<th>Count</th>
<th>% of Codes</th>
<th># of Sites</th>
<th>% of Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sketches</td>
<td>74</td>
<td>0.4</td>
<td>56</td>
<td>23.2</td>
</tr>
<tr>
<td>Graffiti</td>
<td>75</td>
<td>0.4</td>
<td>50</td>
<td>20.7</td>
</tr>
<tr>
<td>Other</td>
<td>184</td>
<td>1.1</td>
<td>43</td>
<td>18.3</td>
</tr>
<tr>
<td>CommercialDesign</td>
<td>63</td>
<td>0.3</td>
<td>43</td>
<td>17.8</td>
</tr>
<tr>
<td>StreetArt</td>
<td>38</td>
<td>0.2</td>
<td>35</td>
<td>14.5</td>
</tr>
<tr>
<td>Murals</td>
<td>39</td>
<td>0.2</td>
<td>32</td>
<td>13.3</td>
</tr>
<tr>
<td>Tags</td>
<td>22</td>
<td>0.1</td>
<td>17</td>
<td>7.1</td>
</tr>
<tr>
<td>3D</td>
<td>18</td>
<td>0.1</td>
<td>16</td>
<td>6.6</td>
</tr>
<tr>
<td>Characters</td>
<td>59</td>
<td>0.3</td>
<td>15</td>
<td>6.2</td>
</tr>
<tr>
<td>Pieces</td>
<td>37</td>
<td>0.2</td>
<td>15</td>
<td>6.2</td>
</tr>
<tr>
<td>Stencils</td>
<td>20</td>
<td>0.1</td>
<td>13</td>
<td>5.4</td>
</tr>
<tr>
<td>Bombs</td>
<td>13</td>
<td>0.1</td>
<td>12</td>
<td>5.0</td>
</tr>
<tr>
<td>Throwups</td>
<td>16</td>
<td>0.1</td>
<td>12</td>
<td>5.0</td>
</tr>
<tr>
<td>Letters</td>
<td>19</td>
<td>0.1</td>
<td>10</td>
<td>4.1</td>
</tr>
<tr>
<td>Productions</td>
<td>12</td>
<td>0.1</td>
<td>10</td>
<td>4.1</td>
</tr>
<tr>
<td>Stickers</td>
<td>14</td>
<td>0.1</td>
<td>10</td>
<td>4.1</td>
</tr>
<tr>
<td>Digital</td>
<td>8</td>
<td>0.0</td>
<td>8</td>
<td>3.3</td>
</tr>
<tr>
<td>TrainWholecars</td>
<td>13</td>
<td>0.1</td>
<td>8</td>
<td>3.3</td>
</tr>
</tbody>
</table>
The last category of work-related codes is the Location codes. Location can be considered a common attribute to document for most traditional artworks, but it holds special significance in the graffiti art community. Graffiti art styles are passed on from older, more established writers to younger ones, and graffiti writers will often “bite” or copy work they admire by others. Styles can be associated with geographic locations around the world as well as with individual artists. Having as precise a location as possible for an individual work is desirable for those wishing to see the work in person as well as for those who research the art style and its evolution across time and space. The value of location information is counterbalanced by the desire of artists acting illegally to remain anonymous and not leave a trail by which they can be tracked by law enforcement. This tension is evident in the very consistent lack of precise location information available across all 241 websites.

The most commonly employed level of geographic location information was by city, followed closely by country. Only one site got close enough to mention a street address, while another once mentioned an intersection of streets. Thirteen sites referenced location via specific landmarks that might be recognizable to some familiar with the next level up in the location hierarchy, such as city name. Parts of cities were also used by thirteen sites. This would include mention of a specific borough of New York City, or a cardinal direction employed with a city name, such as East L.A. Many sites employed numerous levels of geographic faceting, starting by continents or countries, and working down through specific states and cities. A number of websites were geographically focused and indicated works from countries outside their focus with a gallery for World graffiti, a type of “other” code.

Table 4. Location codes and their usage across all sites.

<table>
<thead>
<tr>
<th>Location Codes</th>
<th>Count</th>
<th>% of Codes</th>
<th># of Sites</th>
<th>% of Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cities</td>
<td>1637</td>
<td>8.6</td>
<td>43</td>
<td>17.8</td>
</tr>
<tr>
<td>Countries</td>
<td>543</td>
<td>2.8</td>
<td>37</td>
<td>15.8</td>
</tr>
<tr>
<td>SpecificLandmarks</td>
<td>73</td>
<td>0.4</td>
<td>13</td>
<td>5.8</td>
</tr>
<tr>
<td>CityParts</td>
<td>94</td>
<td>0.5</td>
<td>13</td>
<td>5.4</td>
</tr>
<tr>
<td>World</td>
<td>22</td>
<td>0.1</td>
<td>12</td>
<td>5.0</td>
</tr>
</tbody>
</table>
4.0 Conclusion and Further Research

This research has analyzed the organizational facets employed across a large number of websites that share graffiti art images and how they are further broken down. The work is the first of its kind to describe the results of efforts by a distributed group of image collection managers to organize graffiti images online. It is valuable in that it provides insight into the facets of organization in use around the world to categorize artworks not often collected or documented by traditional art or cultural heritage institutions. The methodology employed has been fruitful for description of current organizational practice.

Several factors intrinsic to the art form complicate the ability to effectively use traditional methods of documentation for works of art, including in large part the often illegal nature of creating graffiti and street art. Legal issues often contribute to the obfuscation of common aspects of art documentation, such as where works are, the identities of those who created them, and dates for creation, change, and destruction of works. Commonalities in practice are easily seen from this research, but there is a lot more that can be gained by further study. Knowing how these works are organized right now is only half of the equation that could lead to development of systems to serve not only those who maintain these diverse collections, but the many users who approach them as well. Further research into the needs and desires of such collection users could fill in other missing facets that may be extremely useful.

While conducting this research, it was found that a number of the websites studied were very well developed. Nineteen of the 241 sites had very large collections, employed notably consistent, very granular use of facets, were active and adding new images, and included clear and in-depth information about the sites themselves. These 19 sites have been noted for further study. Much more analysis was carried out on the websites and interviews of website curators were also conducted, adding still more valuable information about why certain facets and terminology are used. These additional details, in combination with study from the user perspective, would add to the growing amount of information that could be used to design systems for the documentation and organization of graffiti art and street art, as well as other types of found art and ephemera.

References


Schacter, Rafael. 2014. “The Ugly Truth: Street Art, Graffiti and the Creative City.” Art & the Public Sphere 3, no 2: 161-76.


Understanding Personal Online Risk to Individuals via Ontology Development

Abstract:
This paper describes the development of an ontology of risk as a way of better understanding the nature of the potential harms individuals are exposed to when they disclose personal data online. The ontology was designed to be compatible with BFO, the Basic Formal Ontology, which is intended to promote interoperability. Ontologies from domains such as genetics and medical research are in many instances designed to conform to BFO. An initial exercise to monitor the online activity of six participants from the library and information services community helped to identify the points at which personal data is disclosed during online activity. It also explored the motivations for these disclosures, by questioning participants about their perceptions of risk. The resulting analysis suggested that an ontology would be better than a typology to represent the complex relationships between risk concepts. Terms were also extracted from existing terminologies. Risk scenarios were developed and tested during a formative workshop and incorporated into the ontology. A potential application of the ontology is to identify clusters of risk and map the factors that contribute to specific risks.

1.0 Introduction
This research arose from an investigation into the nature of the risks associated with online disclosure of personal information. Interactions with online systems and social media platforms use an economic model based on the sale of personal data (Enders et al. 2008). For instance online behavioural advertising has been a remarkably effective model that has led to the growth of companies such as Facebook, which was able to announce profits of $18.5 billion on revenue of $70.7 billion in 2019 (Facebook 2020).

In return for disclosing personal data, individuals gain ‘free’ access to online services. When faced with risk, feelings should be considered alongside rational decision making (Loewenstein et al. 2001; Finucane and Holup 2006). Behaviour models tend to emphasise conscious, rational decision-making during online transactions involving personal data (Kehr et al. 2015). This has been characterised by many researchers as the ‘privacy calculus’. The perceived benefits are judged to outweigh the perceived risks of disclosure.

Individual risk and public safety are a focus for current UK government policy (DCMS 2019). In the European Union privacy concerns have been reflected in the General Data Protection Regulation (GDPR) (European Parliament 2016).

The purpose of this research is to understand the nature of the risks faced by individuals when they conduct online transactions. The description and categorizing of risks may help with the delivery of more effective mechanisms for managing those risks.

Baldwin et al (2010) argue that the purpose of regulation is to manage risk. Although legislation is the primary means of regulation adopted by government, it is not the full picture. Lessig (2006) encapsulated one aspect of internet regulation by the phrase “Code is Law”. The way in which systems are designed affects the way in which they operate. Cavoukian (2012) extended this idea with the concept of ‘privacy by design’. Haynes et al (2016) go on to suggest that a number of regulatory mechanisms (coding, self-regulation, market response and law) work in concert to regulate access to personal data on social networks. Mapping the risks and their relationship with causes and effects may produce better insights into effective responses to this public safety issue.
This research sets out to examine the nature of the risks faced by individuals when they engage in online activity. The research considers the following questions:

- What is the nature of the risks that individuals face when using the internet?
- Is there an existing typology of online risk?
- Can an ontology of risk be developed to represent risk relationships more effectively than previous typologies of risk?

2.0 Literature review

2.1 Nature of online risk

Risk is an elusive concept, the definition of which depends on the context (Fischhoff, Watson, and Hope 1984). Aven and Renn (2009, 2) define risk in the following terms:

A. Risk is expressed by means of probabilities and expected values
B. Risk is expressed through events/consequences and uncertainties

Simply put, risk is the “effect of uncertainty on objectives” (ISO 2009, 1).

Risk applies to individuals, organizations, governments and societies. When considering the risk to individuals it is necessary to make a distinction between risks to personal privacy and risks associated with disclosing personal data (e.g. via data breaches, as well as voluntary disclosure). The privacy calculus captures the concept of perceived individual risk as well as benefits associated with disclosure of personal data (Dinev and Hart 2006). Studies have found that there is an inverse correlation between severity of perceived risks and willingness to disclose personal data (Dinev and Hart 2006).

Some studies have described the apparently paradoxical result where individuals disclose personal data despite perceived dangers associated with doing so (Gimpel, Kleindienst, and Waldmann 2018). Privacy paradox studies tend to depend on interviews with individuals about what they would do in hypothetical situations (Gimpel, Kleindienst, and Waldmann 2018; Min and Kim 2015). Work by Acquisti and Grossklags (2005) suggested that there is a discrepancy between intention and actual behaviour.

2.2 Using an ontology to describe risk

This research initially set out to develop a taxonomy of risk based on harm to individuals. This would allow hierarchical relationships between concepts. Entities in a taxonomy can be grouped by common origin (phylogeny) or by similarity (morphology) (Gnoli 2017).

Solove (2006) provides a classification of harms, which is a starting point for categorizing risks. These largely predate the advent of social media and need to be updated to incorporate the spectrum of online harassment which can range from bullying through to hate speech. Skinner, Song, and Chang (2006) developed a taxonomy of risk based on three dimensions or views: time, space and matter. This was specifically developed in the context of collaborative environments and needs validation with empirical data. Wright and Raab (2014, 290–91) identify examples of harms based on privacy principles. These both feed into an initial identification of online harms. Haynes and Robinson (2015) set these risks in a network of interconnected risks and consequences.
2.3 Complexity of relationships and ontologies

The decision to use an ontology was based on the ability to define classes of concept and to describe different types of relationship between those classes. Ontology development has been extensive in the biomedical area and this provides a corpus of experience that can be applied elsewhere. Some attention has been paid to other domains such as project management, business processes and cyber security, either using ontologies as a tool for risk assessment (McKone and Feng 2015; Mohammad et al. 2015) or as a means of mapping the relationships between different elements of risk and specific instances of risk events. Perhaps the most directly relevant work is the review of ontologies covering cyber risk which seemed to emphasise vulnerabilities and exploitation by an attacker. There was less emphasis on the concepts of likelihood and impact, which were included in only 3 of the 10 ontologies reviewed by Oltramari and Kott (2018). The authors highlight the problem of estimating probabilities and impact levels in a dynamic environment where the behaviour of a target affects the outcomes. So, for instance if a targeted organization improves its security measures, a potential attacker will switch their attention to another, more vulnerable target. They also speculate that it is impossible to determine the outcomes without knowing more about the motivation of the attackers.

An ontology of online risk needs to reflect the complex nature of risk and the need to incorporate concepts such as: Vulnerability, Threat, Incident, Consequence, Harm and Response. Some of these classes also have properties that are defined in their schemas. For example, it might be useful to incorporate the idea of impact of a Harm or the probability of an Incident into the description of a risk scenario.

3.0 Methods
3.1 Creation of the ontology

Early prototyping used the Graphite system provided via the Synaptica interface. This was intuitive and allowed experimentation with different data formats and development of schemas. This development environment allows export into an OWL-compatible system so that it can plug into high-level ontologies such as the Basic Formal Ontology (BFO).

The ontology development was based on the approach described by Arp et al (2015), who describe four general principles of ontology design:

1. Realism – an ontology is a representation of reality, which is supported by evidence and observation
2. Perspectivalism – reality is too complex to be represented by a single approach. Ontologies should therefore aim to be relevant and accurate within a specified domain
3. Fallibilism – an ontology will change as our understanding and knowledge of a domain develops. It is therefore necessary to be able to keep track of different versions of an ontology and the changes made
4. Adequatism – room must be made for all the types of entity that exist within the domain of the ontology

Arp et al. (2015, 44) suggest that ontologies are representations of reality rather than models of reality based on mental concepts:
Realism in ontology is based further on the idea that with the aid of science we can come to know the general features of reality in the form of universals and the relations between them. This realist approach has a number of general consequences. First, it implies that ontologies are representations of reality, not of people’s concepts or mental representations or uses of language. This presents some real challenges in dealing with human behaviour and motivations. When looking at privacy this research is concerned with motivations to disclose personal data online and the harms (and benefits) that might result. The harms themselves may depend on the perceptions of the individual, so that similar events might be viewed very differently by different individuals.

What is the ‘reality’ we are trying to represent with this ontology? The fact of people’s perceptions is a reality that is captured in attitudinal surveys. They provide a snapshot if what people thought at a particular point in time – and of course they may change in light of experience, better understanding of online harms or education about privacy risks.

Risk can be seen as part of the ontology of social reality rather than objective reality, because it depends on agency: “risk belongs to this subjective ontology [of social reality]. Thus, risks are real, but only insofar as there is a social reality in which subjects engage in risk taking.” (Merkelsen 2011, 894).

3.2 Choice of software

The ontology was designed to be hospitable to RDF data to allow for import from other ontologies and export of the resulting ontology to new environments. The Protégé system developed at Stanford was considered as a suitable platform because it is widely used and has an active community of developers. It supports OWL, which is a W3C standard. The Synaptica Graphite system was also considered for this exercise and was eventually selected because of its terminology management features and the support available to the researcher.

3.3 Development of the ontology

The methodology for development of the ontology was described in a previous paper (Haynes 2019). Noy and McGuiness’ (2001) iterative approach was adopted and applied to the seven-step method for ontology development of Arp et al (2015).

3.4 Testing and validation

The ontology design was tested in a workshop with 14 researchers and practitioners with backgrounds in: knowledge organization, information governance, cybersecurity and information science. Participants worked in groups to examine the proposed representation of risk and to provide critiques to refine it.

An initial set of risk incident types was incorporated into the ontology as a set of scenarios, based on standard definitions and on descriptions in the literature. A degree of normalisation was required for consistency. Seminar participants were asked to explore risk scenarios to identify the consequents and harms that could result from each type of incident. They were also asked to consider the causes that contributed to the incident. The responses were consolidated and expressed as relationships, which entered into the ontology. The relationship network was then explored and graphs generated to illustrate the connection between different entities in the ontology.
3.5 Visualization of graphs

The graphs representing the relationships were shown using the visualization tool within the Synaptica Graphite system. This is an interactive system that allows exploration of the relationship between nodes and navigation through the landscape of risks, their causes and consequences.

4.0 Results

4.1 Scope of the ontology

The scope of the ontology was defined during the early stage of the project and was based on the overall objective of better understanding risk to individuals. The scope of the ontology is described more fully in Haynes (2019, 171–72) and can be summarised as follows:

The ontology covers online hazards faced by online users and the resulting consequences and harms to the individual. It shows the cause and effect relationships between threats, incidents and consequences of disclosing personal data online. The main purpose of the ontology is to map different types of hazards that individuals face and the possible mitigating actions that they could take. It will also identify similarities between different hazards and to identify ways in which they might be addressed.

4.2 Evolution of the representation of risk in the light of feedback

During the workshop, the initial representation was endorsed with some modifications to align it more closely with the cybersecurity view of risk rather than the project management view. Figure 1 shows the revised representation of risk, which incorporates feedback from the formative workshop. Risk is now defined in terms of threats that exploit vulnerabilities in systems. The threats could be malicious or accidental. Risk events are classed as Incidents. As well as mitigating actions to lessen the impact of an incident, there are also avoiding actions and defending actions to reduce the likelihood of an incident and to reduce or eliminate the threat and/or vulnerability of a system.

![Figure 1: Modified Representation of Risk](image)

There was some discussion about whether consequence and harm should be separated. Examination of instances of this representation suggest that it is useful to distinguish...
between the consequence of an incident and the harm to an individual. For example, during a data breach incident, personal bank account details might fall into the hands of criminals and the harm to the individual might be loss of money. The harm is not necessarily realised because the bank may take mitigating action, or the criminals might fail to exploit the data.

4.3 Scenarios
A set of scenarios was developed from reports in the literature, the case studies conducted with the volunteers and development of scenarios during the workshop. Table 1 lists the scenarios used to test different types of risk faced by individuals. They were used to explore the relationships between the causes of a risk and its consequences and these were captured in the ontology.

Table 1 - Scenarios used to develop the ontology

<table>
<thead>
<tr>
<th>Risk</th>
<th>Incident scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLICK-BAIT</td>
<td>Fall down a click-bait rabbit hole</td>
</tr>
<tr>
<td>CLOUD STORAGE</td>
<td>Data breach of cloud documents</td>
</tr>
<tr>
<td>DIGITAL ASSISTANTS</td>
<td>Digital assistant self-launches</td>
</tr>
<tr>
<td>ILLEGAL SITE</td>
<td>Visit an illegal site</td>
</tr>
<tr>
<td>LOCATION TRACKING</td>
<td>Location tracking made public</td>
</tr>
<tr>
<td>NON-SECURE SITE</td>
<td>Land on non-https site</td>
</tr>
<tr>
<td>ONLINE BANKING</td>
<td>Bank login details revealed</td>
</tr>
<tr>
<td>ONLINE PURCHASES</td>
<td>Data breach of online purchase transaction</td>
</tr>
<tr>
<td>OUT-OF-DATE SOFTWARE</td>
<td>Use out of date software</td>
</tr>
<tr>
<td>PHISHING</td>
<td>Respond to phishing email</td>
</tr>
<tr>
<td>PICTURES ON SOCIAL MEDIA</td>
<td>Hostile response to photo posted on social media</td>
</tr>
<tr>
<td>PROFESSIONAL NETWORKS</td>
<td>Employer discovers job-seeking activity</td>
</tr>
<tr>
<td>RE-USE OF PASSWORDS</td>
<td>Re-used password is detected</td>
</tr>
</tbody>
</table>

4.4 Exploring the network of relationships
The modified representation of risk is based on different relationships between the concept classes. Table 2 shows the classes and their relationships within the ontology. Many of these relationships have reciprocals. So for instance, the top term ‘Psychological harm’ in the ontology scheme Harm, has narrower terms: ‘Annoyance’, ‘Fear’ and ‘Worry’. Each of these has a reciprocal broader term relationship with ‘Psychological harm’.

Table 2 - Relationships allowed between concepts in different classes

<table>
<thead>
<tr>
<th>Subject (class)</th>
<th>Predicate(s)</th>
<th>Object (class or property)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consequence</td>
<td>broader/narrower</td>
<td>Consequence</td>
</tr>
<tr>
<td>Consequence</td>
<td>leadsTo</td>
<td>Consequence</td>
</tr>
<tr>
<td>Consequence</td>
<td>leadsTo</td>
<td>Harm</td>
</tr>
<tr>
<td>Harm</td>
<td>broader/narrower</td>
<td>Harm</td>
</tr>
<tr>
<td>Harm</td>
<td>hasProperty</td>
<td>Impact</td>
</tr>
<tr>
<td>Incident</td>
<td>broader/narrower</td>
<td>Incident</td>
</tr>
</tbody>
</table>
The visualization of the ontology demonstrates the complex relationships between concepts (Figure 2). For instance ‘breach of cloud storage’ is a scenario in the Incident scheme. It is a consequence of ‘use of cloud services’ (a prerequisite in event tree analysis) and/or ‘data theft’. It leads to ‘loss of confidentiality’ and ‘consequential loss’.

From Figure 2 we can see that Use of Cloud Services is a vulnerability and that has a number of subordinate relationships. Data theft on the other hand is classed as a threat because it implies intent on the part of an agent.
Going the other way, a breach of cloud storage could lead to loss of confidentiality, which in turn could lead to loss of reputation (a harm). Loss of confidentiality could also result from the self-launch of a digital assistant. There are likely to be other incidents that could lead to this consequence. The breach could also lead to a consequential loss resulting in financial loss to an individual, another harm. Some relationships are two way. For instance, ‘Breach of cloud storage’ could be both a cause and a consequence of ‘Loss of confidentiality’. This illustrates the greater richness of description that is possible using an ontology rather than a taxonomy.

4.5 Inferences from the ontology

As well as providing a helpful visual display of the relationships between different aspects of risk, the ontology allows navigation and exploration of different aspects of risk. This could be valuable in tracking relationships and identifying connections that are not immediately obvious on initial inspection. A possible development of this research would be to consider the eigenvector values of each node to determine closeness and identify potential clusters of concepts (Hansen, Schneiderman, and Smith 2011). This may reveal deeper structure in the set of scenarios in the ontology.

5.0 Discussion and conclusion

5.1. Addressing the research questions

The project set out to explore the nature of risks that individuals face when using the internet. One way of doing this is to develop a taxonomy. Taxonomies are based on hierarchical relationships and do not allow for the complex relationships between risk concepts. For this reason, an ontology was developed instead. It was based on pre-existing work as well as industry definitions of vulnerability and threats (Haynes 2019). However, these definitions tended to be focused on technical issues and consequences to systems, or organizations. This ontology shifts the focus on to people and the impact of online incidents on individuals. It explores by means of scenarios the relationships between Vulnerabilities, Threats and Incidents and then the outcomes of Incidents in terms of Consequences and Harms. The ontology also includes Responses that could mitigate these risks.

5.2 Limitations

The analysis of scenarios is based on one researcher’s interpretation of data gathered from a small group of experts. To some extent this is subjective and needs a more rigorous evaluation – possibly by means of a Delphi study. This would allow a panel of experts to arrive at a consensus about the concepts and relationships associated with the scenarios.

5.3 Future Development

The next stage of development for this ontology is to populate it with instances from a variety of sources, including reports in the press, incident data from data protection regulators and case studies in the literature. This would test how well the scenarios describe the reality of online risks to individuals. It would also provide the groundwork for creation of linked data sets, which could be analysed to inform policy on online safety.
Acknowledgement
This research was supported by the Royal Academy of Engineering and the Office of the Chief Science Adviser for National Security under the UK Intelligence Community Postdoctoral Fellowship Programme (Grant No. ICRF1718/154). The Graphite system used to develop the ontology was provided by Synaptica Ltd. The research was conducted during Dr Haynes’ Fellowship at the Department of Library and Information Science at City, University of London. Thanks to colleagues at Napier and the anonymous reviewers for their valuable comments and suggestions.

References


The Use of Community to Organize Knowledge
The Case of an Energy Company

Abstract:
This research conducted within an energy sector company brings together both information systems and knowledge organization (KO). It is based on a case study, aiming to analyze the ‘interface’ question in a comprehensive way through the building of an information system dedicated to the organization of knowledge within a community of practice. Through this case, we will develop an approach related to KO technologies that highlight the importance of the 'interface' not only as user interface in software but, moreover, as a pathway between users, communities and organizations necessary for promoting a common understanding and use of knowledge in an efficient way.

1.0 Introduction
In this political and economic understanding of the society that is the ‘Information Society’ or the ‘Knowledge Society’ (first used by Drucker (1969) then by the UNESCO), the question of how to organize the knowledge is essential. From decades knowledge organization systems (KOS) (Mazzocchi 2018) such as classifications, and documentary languages, topic maps, ontologies (more recently through the use of information technology), were developed in order to optimize the way we use knowledge.

The organization of knowledge can be thought of through the prism of digital technology because the theoretical framework of knowledge organization (KO) allows us to ‘develop methods to guide effective practices to exploit our knowledge in a digital environment’ (Beau 2012, 1). The knowledge organization system in this way, ‘constitutes a common language, either for the design of an information system, or, more generally, for sharing knowledge concerned by different carriers. It can be used as a framework to express knowledge shaping in the field in as comprehensive and complete manner as possible’ (Mahé et al. 2010, 66). The objective of these KOSs is therefore to ‘define principles for describing a domain to facilitate the classification and search for more or less abstract items: documents, persons, places, products, opinions or activities’ (Zacklad and Giboin 2010, 8) and thus facilitate knowledge dissemination. Moreover, in organizations where dematerialized knowledge, or even produced natively in digital formats, has become prevalent (Martínez-Ávila 2015; Fujita and Pinheiro 2016).

The concept of interface can be viewed as a user interface in order to access to knowledge. In KO publications, authors are using generally the term ‘interface’ in this meaning. However, an interface can also be seen as a pathway between groups of people or inside a community (here for instance between science and society (Puente-Rodríguez, Bos and Koerkamp, 2019)). In this research, we will explore the interface aspect, not only as a computer interface but to show how a KOS can serve as an interface at various levels of an organization.

In this paper, we will focus on the interface aspect of knowledge organization as follows: we first, develop our literature review in order to define more precisely the research questions and scope. Then we define our methodology, show results and our
findings that are going to be discussed. Finally, we show our search shortcomings, conclusion and limits.

2.0 Literature Review

According to Broughton et al. (2005, 133), the concept of KO is mostly about the use of knowledge organization systems (e.g., classification, thesauri, semantic networks) and the process of organizing this knowledge. If KO is mainly related to ‘memory institutions’, many organizations are working now over the way they organize their knowledge. In this article, we will focus over this KO dimension in a private organization to highlight an aspect related to KO: the interface dimension.

The professional context in which the members of an organization operate is then a specific context in terms of knowledge organization. This forces the organization to create its own organizational model and technical systems to best meet its objectives by empowering its employees to act.

Knowledge produced within an organization can be considered as organizational knowledge (Bibikas et al. 2008; Coakes, Coakes, and Rosenberg 2008; Yang, Fang, and Lin 2010) as an entity is able to produce new knowledge regarding its needs and then disseminate it throughout the organization (including in services, products or systems). This production of organizational knowledge is related to the actions of members who are part of the whole (the organization) and create in the context of their mission knowledge to achieve ‘some end’ (Nonaka and Takeuchi 1995, 58). Before becoming organizational knowledge, knowledge is already individual or even communal within a group (Merali 2000; Allard 2004; Kaschig, Maier and Sandow 2016).

The willingness to share knowledge among employees remains fairly recent (in relation to the development of knowledge management (Bell DeTienne et al. 2004)): ‘organizations have only recently begun to expect their employees to consistently share and exchange knowledge; in the past, organizations typically urged workers to pursue individual goals and rewarded them on the basis of individual performance and knowhow’ (Biron and Hanuka 2015, 655) and this with the aim of being competitive (Chen and Fong 2015; Martinez-Gil 2015).

The community has an essential role in the production and management of organizational knowledge by considering that ‘real knowledge management is not possible without true community’ (Hassel 2007, 193), in fact ‘in a strict sense, knowledge is created only by individuals. An organization cannot create knowledge without individuals’ (Nonaka and Takeuchi 1995, 58); from individuals, knowledge can become organizational by being shared in communities of practice and more broadly in the organization. These statements resonate with the consideration that is by allowing, in the communicating organization, to co-construct meaning that ‘emerges within communities and that its analysis should not be dissociated from the social, historical, cultural and political dimensions’ (Lemke 1995, 9 cited by Hachour 2011, 202).

Thus, individual knowledge can become ‘organizational’ by being shared in the communities of practice and more widely in the organization. The professional situation in which the members of the organization operate is in fact a specific context in terms of knowledge organization. For instance, an action carried out by a member of the organization involves a ‘situated action’ (Guyot 2000) and therefore to a situation where the actor is confronted with a need for established knowledge, or even the production of
a new knowledge if the technical situation is a new one. The knowledge produced therefore comes from the professional context in which the need to solve concrete problems appears.

This idea is in line with the fact that ‘the collective competence of actors is based on the existence of networks that ensure knowledge sharing’ (Alter 2000, 267). This situation can be illustrated by Castro Goncalves (2011) who highlights the fact that learning in an organization is supported by interactions between individuals confronted in their tasks. If those researches illustrated the role of community to create knowledge, however, it remains the question of how this community can organize knowledge the way they need it. In many cases, KOS are controlled at the organizational level by experts that are in a ‘for use’ approach (Folcher 2015). These systems are set up, developed and mediated by experts who are sometimes quite far from the operational situation. They, then design these systems in a logic ‘for use’ before being set out for the actors. Instead of developing them in a ‘in use’ (ibid.) approach, in which end-users’ usage patterns are registering in the developed software. This approach is then a source of legitimization for the system and a way of sharing the view of the community regarding the way they organize knowledge toward the rest of the organization.

This literature review leads to the following research question, how the development of a hybrid KOS by its future users can be seen as an interface at various levels in the organization?

3.0 Context of the research and methodology

In order to investigate this question, we choose a case study approach in an organization. The data collected allowed us to design this progressive development of a new KOS. It is a result of a one-year participating observation (Soulé 2007). Other data were collected during workshops aimed at identifying more precisely the expectations of future users. Participating observation (i.e.: daily observations) and workshops (which allow to focus the attention of a group over a specific question, for instance hover the kind of knowledge they used) are complementary in order to adopt an “insider” position close to the protagonists studied. This approach elects us to deal with how the knowledge is socialized, produced and organized, and in this context of the interface question connects situated knowledge within a community and more broadly in the whole organization. In this way, regarding our research question, we can manage to apprehend choices and views of the actors in a more comprehensive way; for instance, taking in consideration political aspect or relationship between individual/direction.

The company ‘Alpha’ is an important company in the energy sector, according to the INSEE nomenclature, which operates throughout France. The population of our study (also called ‘Territorials’ and represents around 45 individuals) is located in the Île-de-France region (IDF) and works in public relations positions. We can describe those ‘Territorials’ as a community of practice (Wenger 1998) since they share codes and routines that are part of the collective’s social practices (Coulon 2002); they have then internalized the values conveyed by this community and its system of representation. Quarterly meetings, gatherings and seminars organized during the year are opportunities for them to meet, exchange and at depth to strengthen their capacity to produce practices. The use of systems is then limited to devices such as e-mail, calls or ‘business’ software, but without any specific space for organizing the knowledge of this community or
facilitating its sharing. As part of their job, in relation with local actors, the knowledge they produce is kept individually by employees or sometimes recorded in textual documents or even in Excel spreadsheets.

4.0 Results

‘Territorials’ are in the logic of ‘actionable knowledge’, i.e. knowledge intended to produce an action and effects (Argyris 2003), that the business software (call @T) they used before does not allow. In the case of @T, its use has never been approved by the ‘Territorials’ but was imposed by the national management as a case tracking tool. Since its implementation in 2012, the ‘Territorials’ have been constantly developing other systems that better meet their needs for information and project monitoring.

The request expressed by the ‘Territorials’ was the construction of an information system allowing creating links between their data, actors, projects and territories in order to have actionable knowledge. Progressively, the participants in the workshops noted that the observations regarding the system could not be carried out independently of those concerning their practices. They even had an official mission letter from a member of their steering committee to pursue this aim. In conjunction with the technical part, the associated approach has been reviewed to identify current practices. Far from simplifying an approach of technical determinism or innovation determinism, the technical system is built through practice and use and goes beyond technical aspects in order to bring about through its mediation action a set of translations that build a sociotechnical ecosystem that has not yet stabilized (Hoareau 2014).

It is on the basis of this observation that the ‘Territorials’ have developed an entity-association or entity-relationship scheme proposed by Chen (1976). This is how the KO dimension emerged in the project. This work on the system is essential, an information system is above all a symbolic system of representation (Bélisle 2002) which is mobilized here and which we had to present/model through the entity-relationship scheme and the vision on their profession so that this view can be implemented in the final product. This is then composed by representations designed and interpreted by the ‘Territorials’ (business classifications, integration of their processes, etc.) and links collective and/or individual actions by a technological base. Regarding their needs and the solution that was developed to answer practically to them, the KOS is then a hybrid one that aggregate: terminology (in order to have a common vocabulary between them), a knowledge base (in which they can preserve their knowledge) and also a semantic network (this semantic network is inspired by linked data, yet it’s simplified by using the graph database NEO4J. This database allows storing elements and create semantic relations between them). The result of the semantic network looks like a topic map in the user interface in order to allow them to ‘navigate’ through their knowledge and their concepts/entities figure N° 1:
All those elements aggregated in the final hybrid KOS are interoperable in order to facilitate the communication between the components and the other information system in the organization.

Regarding the individual, community and organizational level, we can summarize the results at those levels through the Table N° 1 bellow:

Table 1: Comparison between individual, community and organizational level using the information system with hybrid KOS

<table>
<thead>
<tr>
<th>Individual level</th>
<th>Community level</th>
<th>Organizational level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Save time with faster identification of the right interlocutors for a project.</td>
<td>Harmonization of practices and vocabularies between IDF ‘Territorials’.</td>
<td>Implementation of knowledge continuity.</td>
</tr>
<tr>
<td>Simplification for tracking information and fewer various information flows.</td>
<td>Construction of a socio-technical system so as to become a virtual community of practice.</td>
<td>Use of a new technology within the organization: graph-oriented databases.</td>
</tr>
<tr>
<td>Consideration of their requests and recognition of the specificities of their activities.</td>
<td>Work of reflection on their job and the way it is done.</td>
<td>Enhancement of the organization’s information assets in order to make it more efficient.</td>
</tr>
<tr>
<td>Deduce new information or knowledge through graphic modeling which becomes a support to the analysis of ‘Territorial’ data.</td>
<td>Possibility to do more collaborative work. By adding data and information via the online tool, they enrich their collective knowledge heritage which can then be consulted and used by all.</td>
<td>Valuing the members of the organization by showing consideration for their needs or expectations.</td>
</tr>
</tbody>
</table>
Capturing weak signals and setting up inductive logic from the visualization.  
Production of values, practices, benchmarks for community members.  
Development of an interface between various actors with a common need of knowledge.

It’s obvious the system plays an interface role at various levels and results of its actions would depend on this level.

5.0 Discussion

By designing the information system ‘in their uses’ and practices, and by questioning their profession, this seems to be an approach that limits the risk of rejection of a hybrid KOS and makes it closer to the operational reality. From an organizational point of view, the development of a specific information system shared by the ‘Territorials’ raises questions about their extremely individualistic culture (for instance address book is a tool for their work and at the service of their careers) to a more collaborative culture. Moreover, in this case, the creation of a shared vocabulary in the hybrid KOS is an opportunity to reinforce their common culture and to create a bridge between this community and the organization to spread knowledge.

By building their KOS, ‘Territorials’ also make an important analytical work related to their knowledge and their needs of knowledge. To illustrate that, during one workshop arise the question of how to preserve knowledge when someone is leaving? As we know, when a member of the community leaves, the loss of its knowledge for the organization can be detrimental. The implementation of a mechanism to ensure continuity and allow the community to fulfill its mission is then to be considered in a ‘knowledge continuity’ (KC) strategy (Ermine 2010; Biron and Hanuka 2015). It is to alleviate this situation but also to harmonize practices of the members and strengthen the sense of belonging to this group (Ellison, Steinfeld, and Lampe 2007) that an information system was developed on their initiative and with them to transform this community of practice into a fully-fledged virtual community of practice (Tessier, Bourdon, and Kimble 2014).

Thanks to the comprehensive approach allowed by participating observation, we can estimate that this hybrid KOS is then an interface at various levels:

- It is an interface between the members of the community of practice that enhance (first of all) the possibility of sharing knowledge. The development of a common controlled terminology or of a semantic network is then useful to reinforce the community of practice and harmonize practices. Moreover, it’s also an information system that allows them to have a clear-cut idea about their work, their needs of knowledge related to their uses.

- It is an interface between the community and the organization (and therefore other employees). In this specific case, as the community develop its own KOS, they do formalize the way they organize knowledge and how they consider the environment in which they evolve. Furthermore, by using a technological system, it’s an opportunity to reinforce interoperability between their system and the rest of the organization’s information system through the use of API (application programming interface) for instance or the sharing of the used terminology.

- It can at least, be considered as an interface between the community and future members (or new members) in order to facilitate the integration inside the
community by facilitating the access to a shared language and to share knowledge. Then, the KOS enhances knowledge continuity during the time that the community is existing.

This situation underscores the importance of a dualistic evolution of the system and the user, in fact, the user must adapt within the framework of co-evolution where adjustments are made both by the user and to the technical system for optimal operation (Bourguin and Derycke 2005). By authorizing employees to develop their information system, the organization’s hierarchy delegates its ability to control and organize their knowledge.

With the new system, the ‘Territorials’ found themselves facing a situation where they had to produce new explicit knowledge, and, furthermore, must consider the way they organize it in order to use it in an efficient way. The system developed was designed by the future users for collective use, in particular with regard to the knowledge related to their missions.

6.0 Conclusion
This article focused over the study of the building of an information system that is a hybrid KOS (including shared vocabulary, knowledge base and semantic network) by its future users. ‘Territorials’ did an analytical work over knowledge, their knowledge needs and the way they can organize it in order to realize their missions. At that time, this KOS is then an interface between all the ‘Territorials’ that are going to use it and even with the future members of this community.

By building themselves, the aim is to ensure that the KOS is related to their uses. In this configuration, the organization allowed them to create in fact an interface between this community of practice and the rest of the employees.

Throughout this article, we highlight the interface dimension of KO regarding the effect that the development of an information system as a hybrid KOS has over a community of practice and by extension, to an organization. Then, we consider the interface dimension not only as the user interface (what the user is going to see when he is using the software), but also as a pathway of connecting various kinds of users in an organization. It is crucial to highlight this dimension as an essential point step in a KOS design. Moreover, through this example, we also had the opportunity to see how a community is able to design its own KOS and build their relations with the whole company in this way.

7.0 Limitations
At this stage of experimentation, the project is operational on the scale of a single region, but its extension to all regions is envisaged by the national management, which supervise the activity of the ‘Territorials’ if the results on the IDF are convincing. However, as we shall see in the second practical example, regional specificities are sometimes such that there is a strong disparity closely linked to the local context.

References


Philip Hider – Charles Sturt University, Australia

Fiction Genres in Library Catalogues and Social Cataloguing Sites

Abstract
Samples of fiction genres both represented and not represented in the Library of Congress Genre/Form Terms (LCGFT) were compared with respect to their usage in the social cataloguing site, LibraryThing. It was found that the non-LCGFT genres, mostly based on entries in Wikipedia, were markedly more used than were the LCGFT genres. A particular feature of many of the non-LCGFT genres was an element of affect, relatively lacking in the LCGFT sample. It is suggested that there may remain a reluctance in library cataloguing to fully embrace this aspect of genre, and creative works such as fiction, and that this reluctance may be due in part to the traditional, modernist paradigm of the cataloguer as gatekeeper to objects rather than as a facilitator of experiences and feelings that those objects may provide.

1.0 Introduction
The study and description of ‘genre’ has been relatively neglected in knowledge organization (KO), compared with resource attributes such as ‘subject’ (Lee and Zhang 2013). Yet for a wide range of works, including creative works, the concept of genre is central to the way in which they are described, and thus conceptualized, by both creators and consumers. Recently, the Library of Congress developed its Library of Congress Genre/Form Terms for Library and Archival Materials (LCGFT) to help address this gap in library cataloguing (Young and Mandelstam 2013). This paper assesses the extent to which there may still be gaps in LC’s coverage of genre by comparing the use of samples of LCGFT and non-LCGFT fiction genres on the social cataloguing site, LibraryThing. The paper goes onto discuss why missing genres may not have been included in the library vocabulary, and the relationship between professional and commercial genre classifications and the genre folksonomies to be found on social cataloging sites. This discussion draws, in particular, on the critique of the modernist paradigm in library cataloguing and classification by scholars such as Mai (2011), the work of scholars such as Spiteri and Pecoskie (2017) in pointing out the importance of affect in fiction access, and the theoretical framework of art classification originally developed by DiMaggio (1987).

2.0 Literature review
Zhang and Olson (2015) have explored the ways in which genres exhibit qualities of both ‘essences’ and ‘contexts’ in library cataloguing and beyond. On the one hand, genres involve essences that provide stability; on the other they relate to contexts that are fluid. Noting that ‘genre has long been a source of uncertainty and unease in bibliographic control’, the authors consider genre to be ‘an integration of aboutness, of-ness and is-ness’ (Zhang and Olson 2015, 540, 550).

The meaning of specific genres becomes even more complex when it viewed as in an ongoing state of negotiation between different protagonists, such as creators, audiences and intermediaries (Tudor 2012). The influence of different interests and groups on art classification has likewise been included in the theoretical framework originally proposed by DiMaggio (1987), with ‘ritual’ classifications being shaped by commercial, professional and administrative inputs. In the framework, these classifications vary
across the dimensions of differentiation, hierarchy, universality and boundary strength. While the conceptualization of artistic genres has tended to be viewed as more heavily influenced by commercial interests than by the professional views of critics, for example, Brown (2015) notes the exception of ‘feel good’ movies, a category with sometimes positive and sometimes negative connotations that have followed the lead of the film critics rather than the film distributors.

Although genres have, for many decades, played an important role in the everyday description of creative works, their use in library cataloguing and other KO practice has been less than prominent, with far more attention having been given to describing books and other materials in terms of ‘subject’ (Lee and Zhang 2013). Thus there has been a long debate around the concept of ‘aboutness’, but far less attention paid to what it means for a novel, for example, to ‘belong’ to a particular genre category.

Nevertheless, formal controlled vocabularies for the description of genres do exist, with the Library of Congress having developed a separate list, over the past decade, of headings for genres and formats covering a wide range of materials, including fiction. The LCGFT is primarily based on those genres and form subdivisions previously established in the *Library of Congress Subject Headings* (LCSH) (Young and Mandelstam 2013). The LCGFT defines genres as ‘categories of works that are characterized by similar plots, themes, settings, situations, and characters’ (Library of Congress 2018, 3), and this definition has been adopted for the purposes of this paper. For inclusion as LCGFT, terms require ‘literary warrant’, that is, they should be based on items being catalogued, although the terms also need to be justified with reference to authoritative sources (Library of Congress 2018).

Outside of librarianship, another formal classification system that includes fiction genres is BISAC (https://bisg.org/page/BISACEdition), developed for and by the book industry. The classification is not as deep as those typically used in library cataloguing, but is influential in the book world. Because of its commercial roots, its adoption by (some) libraries has been criticized by Martínez-Ávila (2016).

Meanwhile, fiction experts and enthusiasts have the opportunity to contribute to the description of genres on Wikipedia, which divides ‘genre fiction’ into the main genres of 1) crime, 2) fantasy, 3) romance, 4) science fiction, 5) Western, 6) inspirational and 7) horror (https://en.wikipedia.org/wiki/Genre_fiction). The general reading public are also able to ‘tag’ their personal collections with their own genre terms on social cataloguing sites such as LibraryThing (https://www.librarything.com) and Goodreads (https://www.goodreads.com). When aggregated, these terms form a folksonomy (Raftery 2018). Social tagging environments such as LibraryThing have been studied extensively from a KO perspective (e.g. Johansson and Golub 2019; Vaidya and Harinarayana 2016; Voorbij 2012; Bates and Rowley 2011; Lu, Park, and Hu 2010; Adler 2009). Their ‘democratization’ of access provision and KO practice has been championed by scholars such as Mai (2011), who have characterised traditional library cataloguing and classification as modernist and objectivist, with the cataloguer describing and classifying materials according to a single, ‘authoritative’ perspective. Pando and Almeida (2016) have shown how this approach is now being challenged across the field of KO, with a postmodern viewpoint coming to the fore due to the pervasiveness of online technologies that support activities such as social tagging.
While social cataloguing and social bookmarking generate many tags that describe personal relationships with resources (e.g. ‘to read’), many more describe aspects that are, or could be, relevant to other users (Heymann, Paepcke, and Garcia-Molina 2010). Stover (2009) has noted the opportunity the sites provide for the classification of creative works by affect, as well as subject. Spiteri and Pecoskie (2017) have followed up by developing an ‘affect’ taxonomy to cover emotions, tones and association, for use in readers’ advisory services, in this case based on the literature and existing schemes.

Social cataloguing sites with large user communities also provide a good opportunity to gauge ‘user warrant’, defined here as the justification for the inclusion of terms and concepts in an indexing vocabulary on the basis of their likely use by prospective users of that vocabulary. User warrant is often contrasted to other major forms of warrant recognised in KO, including literary warrant, mentioned earlier, expert warrant and cultural warrant (Hider 2015). Svenonious (2000, 135) argues that ‘literary warrant is a necessary but not sufficient basis for admitting terms into the vocabulary of a subject language. This is because there is no guarantee that the vocabulary of those who create the literature of a discipline will match the vocabulary of those who search for it.’ As such, user warrant is a commonly accepted basis for the development of schemes and vocabularies (Hider 2015). Sometimes, user warrant is also distinguished from ‘use warrant’, though other times the terms are used interchangeably (Martínez-Ávila and Budd 2017). In this paper, they will be defined and distinguished operationally, as described in the Methods section. It should also be noted that the literature review did not identify any other study on the social tagging specifically of genres.

3.0 Research questions

This project aims to explore the following research questions:

1) Are there fiction genres with relatively high levels of general use and user warrant which are not covered by the LCGFT?
2) If so, what is the nature of these genres?
3) Broadly, why are libraries not describing fiction in terms of these genres?

4.0 Method

The study devised an index of use and user warrant that could be utilized in the context of a readily accessible social cataloguing site, namely, LibraryThing. A purposive sample of twelve genre terms not included in LCGFT (i.e. as neither preferred nor non-preferred terms) was derived from Wikipedia and, in one case (i.e. ‘pulp’), the Ebay search interface (https://www.ebay.com). The sample was selected on the basis that the terms would generally not be used as subjects, when applied to fiction, as per the distinction made by the Library of Congress between subjects and genres (Young and Mandlerstam 2013). All seven of Wikipedia’s divisions of genre fiction were represented in the sample, as were genres that were described as crossing over different divisions. The twelve genre terms, as well as works cited in Wikipedia and Ebay as examples of the genres, were searched in the Library of Congress catalog (https://catalog.loc.gov/index.html). Literary and collection warrant could be found in all cases. The sample genres were: biopunk, chick lit, dieselpunk, dying Earth, gaslamp, grimdark, hardboiled, inspirational, northwestern, pulp, splatterpunk, and weird. The sample did not allow for
a general comparison between LCGFT and non-LCGFT genre tags in LibraryThing, but it did allow for an exploratory study of possible gaps in LCGFT.

The sample terms were then searched in the LibraryThing interface, which hosts approximately 155M tags, used by roughly 2.3M subscribers (https://www.librarything.com). As such LibraryThing is the world’s biggest social cataloguing site and deemed to be representative of the way in which the public at large describe works of fiction (while acknowledging certain biases, including one toward the English language).

For each sample term, the tags that included either the term exactly or a word-form variant of the term (e.g. ‘biopunk’ and ‘bio-punk’), and that could have been used to represent a fiction genre, were identified. For each of these tags, the number of works for which the tag was used, and the number of subscribers who used the tag, were recorded (LibraryThing collates different printings and editions of the same work, albeit imperfectly). This information was provided directly in the LibraryThing interface.

Tags with variant terms were grouped together. As many tags have been used by thousands of different subscribers for thousands of different works, for the purposes of this study those groups of tags used by fewer than ten subscribers altogether were discarded at this stage. Many of these tags were personal in nature or contained typographical errors.

The total numbers of works and users for the remaining tag groups were then adjusted in some cases for non-genre use, that is, where tags had been used in senses other than a fiction genre. For example, ‘biopunk’ was used as a tag for a book of poetry, while ‘inspirational’ had been used to describe non-literary works that were ‘inspirational’. These adjustments were based on samples of the ten works for which tags had been most used. The final estimates of works and users were then added up for each of the twelve sample genres, providing an index of use warrant and user warrant respectively.

A sample of the 20 LCGFT genres under the ‘Fiction’ heading was selected for comparison with the non-LCGFT sample. As the latter are subgenres, those LCGFT that represented the main genres listed on Wikipedia were excluded. Also excluded were those LCGFT that did not include the word ‘fiction’, as these tended to be ‘forms’, like ‘short stories’, which fell outside of the LC definition for genres above, or in a few cases were non-English terms for non-English language material. Further, those LCGFT listed at more than one level below the ‘Fiction’ heading were not used. The selection of the remaining LCGFT genres was done randomly.

For each of the sample LCGFT genres the process as described above was repeated, except that in this case all of their non-preferred, as well as preferred, terms, as listed in LCGFT, were included as variants to be searched in the LibraryThing interface. The estimated numbers of works and users for the non-LCGFT and LCGFT genres were then compared.

Finally, the most tagged works for each of the twelve non-LCGFT genres were then searched in the OCLC WorldCat database (https://www.worldcat.org), which comprises bibliographic records used in vast numbers of library catalogues around the world. For the top ten works found in WorldCat, the subject and genre headings used in each of their most prominent edition were recorded and analysed for semantic overlap with the sample genre. The headings were also analysed for their overlap across the ten works.
5.0 Findings
The twelve non-LCGFT genres were all represented with tags in LibraryThing. Six of the genres (biopunk, dieselpunk, dying Earth, grimdark, northwestern and splatterpunk) were represented by just one group of variant tags used by at least ten users, the other six by several tag groups with ten or more users; the genre with most tag groups was chick lit, with 15 groups. For some genres with multiple tag groups, however, one or two of the groups were far more used than others. Generally, the most used terms were those that included just the genre term itself (e.g. ‘hardboiled’), but the term qualified with ‘fiction’ or the parent genre(s) (e.g. ‘hardboiled detective’ or ‘hardboiled fiction’), were in some cases also quite often used.

Of the 71 tag groups with ten or more users, across the whole sample of non-LCGFT genres, a majority (42/62 = 68%) appeared to be used virtually exclusively to represent the fiction genre, but about a third (20/62 = 32%) were also used for other concepts to varying degrees. Not surprisingly, the two genres with the tags that exhibited the most mixed use were those with terms that had other common meanings, namely ‘inspirational’ and ‘weird’. The four other genres with polysemic terms were gaslamp, pulp, northwestern and splatterpunk. In some of these cases, the different meaning still related to the fiction genre, but covered the genre more broadly or another literary form (e.g. poetry).

One might perhaps have expected a larger number of tag groups per LCGFT genre than per non-LCGFT genre, as the non-preferred LCGFT terms were also searched in LibraryThing. However, whereas the twelve non-LCGFT genres yielded 71 tag groups, the 20 LCGFT genres yielded just 69 tag groups. Apart from the possibility of different distributions of degrees of ambiguity across these groups, a likely explanation is simply that the LCGFT genres were used less overall in LibraryThing than were the non-LCGFT genres. As with the non-LCGFT genres, some of the LCGFT genres yielded just the one tag group, others yielded several; likewise, while a majority of the LCGFT tag groups were virtually exclusively used for the fiction genre, some (18/69 = 26%) were also used for other concepts, related or otherwise.

Tables 1 and 2 list the estimates for the number of works in LibraryThing that were tagged for each of the non-LCGFT and LCGFT genres respectively, at the time of the study. We can clearly see that overall the non-LCGFT sample was used for many more works than was the LCGFT sample. Of the six genres with tags to more than 10,000 works, four were non-LCGFT, despite the smaller sample size. The top genre was chick lit, followed by magic realist fiction and pulp. At the other end of the scale, there were five LCGFT genres (20% of the sample) with fewer works than the lowest non-LCGFT genre had. If the median number for the LCGFT sample (162) were deemed the threshold for ‘use warrant’, then all bar two of the non-LCGFT genres (dieselpunk and northwestern) would have a strong case for inclusion in LCGFT.

Tables 3 and 4 list the estimates for the number of users in LibraryThing that had tagged for each of the non-LCGFT and LCGFT genres respectively, at the time of the study. The distributions are similar to those of the numbers of works (above), and accordingly we can clearly see that overall the non-LCGFT genres had been used by a lot more LibraryThing members than had the LCGFT genres. Again, chick lit comes out on top by a large margin. Of the seven genres with over 10,000 taggers, five are from the smaller non-LCGFT sample. If the median number of taggers for the LCGFT sample
were deemed the threshold for ‘user warrant’, then again all bar two of the non-LCGFT genres would have a strong case for inclusion in LCGFT. One should also bear in mind that there may well be synonyms for the non-LCGFT terms that were used by other taggers, and for other works, which would increase the non-LCGFT numbers.

Table 1. Works with non-LCGFT fiction genre tags

<table>
<thead>
<tr>
<th>Genre</th>
<th>Works</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chick Lit</td>
<td>74781</td>
</tr>
<tr>
<td>Pulp</td>
<td>27310</td>
</tr>
<tr>
<td>Inspirational</td>
<td>21937</td>
</tr>
<tr>
<td>Weird</td>
<td>14414</td>
</tr>
<tr>
<td>Hardboiled</td>
<td>8179</td>
</tr>
<tr>
<td>Gaslamp</td>
<td>751</td>
</tr>
<tr>
<td>Grimdark</td>
<td>652</td>
</tr>
<tr>
<td>Dying Earth</td>
<td>646</td>
</tr>
<tr>
<td>Splatterpunk</td>
<td>225</td>
</tr>
<tr>
<td>Biopunk</td>
<td>200</td>
</tr>
<tr>
<td>Dieselpunk</td>
<td>139</td>
</tr>
<tr>
<td>Northwestern</td>
<td>52</td>
</tr>
</tbody>
</table>

Table 2. Works with LCGFT fiction genre tags

<table>
<thead>
<tr>
<th>Genre</th>
<th>Works</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magic realist fiction</td>
<td>35345</td>
</tr>
<tr>
<td>Legal fiction (Literature)</td>
<td>10157</td>
</tr>
<tr>
<td>Campus fiction</td>
<td>949</td>
</tr>
<tr>
<td>Philosophical fiction</td>
<td>876</td>
</tr>
<tr>
<td>Utopian fiction</td>
<td>829</td>
</tr>
<tr>
<td>Didactic fiction</td>
<td>488</td>
</tr>
<tr>
<td>Road fiction</td>
<td>371</td>
</tr>
<tr>
<td>Mythological fiction</td>
<td>352</td>
</tr>
<tr>
<td>Social problem fiction</td>
<td>344</td>
</tr>
<tr>
<td>Picaresque fiction</td>
<td>171</td>
</tr>
<tr>
<td>Martial arts fiction</td>
<td>153</td>
</tr>
<tr>
<td>Easter fiction</td>
<td>102</td>
</tr>
<tr>
<td>Prison fiction</td>
<td>73</td>
</tr>
<tr>
<td>Pastoral fiction</td>
<td>62</td>
</tr>
<tr>
<td>Bisexual fiction</td>
<td>55</td>
</tr>
<tr>
<td>Fishing fiction</td>
<td>34</td>
</tr>
<tr>
<td>Transgender fiction</td>
<td>32</td>
</tr>
<tr>
<td>Hunting fiction</td>
<td>24</td>
</tr>
<tr>
<td>Samurai fiction</td>
<td>24</td>
</tr>
<tr>
<td>Nonsense fiction</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3. Users tagging for non-LCGFT fiction genres

<table>
<thead>
<tr>
<th>Genre</th>
<th>Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chick Lit</td>
<td>81056</td>
</tr>
<tr>
<td>Pulp</td>
<td>27639</td>
</tr>
<tr>
<td>Inspirational</td>
<td>21990</td>
</tr>
<tr>
<td>Weird</td>
<td>14676</td>
</tr>
<tr>
<td>Hardboiled</td>
<td>10038</td>
</tr>
<tr>
<td>Gaslamp</td>
<td>776</td>
</tr>
<tr>
<td>Dying Earth</td>
<td>707</td>
</tr>
<tr>
<td>Grimdark</td>
<td>654</td>
</tr>
<tr>
<td>Splatterpunk</td>
<td>230</td>
</tr>
<tr>
<td>Biopunk</td>
<td>202</td>
</tr>
<tr>
<td>Dieselpunk</td>
<td>138</td>
</tr>
<tr>
<td>Northwestern</td>
<td>52</td>
</tr>
</tbody>
</table>
Table 4. Users tagging for LCGFT fiction genres

<table>
<thead>
<tr>
<th>Genre</th>
<th>Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magic realist fiction</td>
<td>37384</td>
</tr>
<tr>
<td>Legal fiction (Literature)</td>
<td>11054</td>
</tr>
<tr>
<td>Utopian fiction</td>
<td>1335</td>
</tr>
<tr>
<td>Campus fiction</td>
<td>1050</td>
</tr>
<tr>
<td>Philosophical fiction</td>
<td>1003</td>
</tr>
<tr>
<td>Didactic fiction</td>
<td>490</td>
</tr>
<tr>
<td>Social problem fiction</td>
<td>445</td>
</tr>
<tr>
<td>Picaresque fiction</td>
<td>264</td>
</tr>
<tr>
<td>Mythological fiction</td>
<td>250</td>
</tr>
<tr>
<td>Road fiction</td>
<td>211</td>
</tr>
<tr>
<td>Easter fiction</td>
<td>108</td>
</tr>
<tr>
<td>Pastoral fiction</td>
<td>102</td>
</tr>
<tr>
<td>Samurai fiction</td>
<td>49</td>
</tr>
<tr>
<td>Prison fiction</td>
<td>31</td>
</tr>
<tr>
<td>Martial arts fiction</td>
<td>22</td>
</tr>
<tr>
<td>Fishing fiction</td>
<td>19</td>
</tr>
<tr>
<td>Bisexual fiction</td>
<td>18</td>
</tr>
<tr>
<td>Hunting fiction</td>
<td>14</td>
</tr>
<tr>
<td>Transgender fiction</td>
<td>11</td>
</tr>
<tr>
<td>Nonsense fiction</td>
<td>0</td>
</tr>
</tbody>
</table>

The distributions in these tables are notable for several other reasons. First, there appears to be hardly any use or user warrant for nonsense fiction, despite its inclusion in LCGFT (or if there is, LibraryThing taggers use a term not covered by LCGFT). Second, chick lit appears to have more use and user warrant than the 20 LCGFT genres put together, demonstrating how a well-used professional vocabulary can nevertheless be hugely at odds with ‘folk’ perspectives on particular works. Third, the exponential-like distributions shows the extent to which genres, or at least fiction genres, can withstand relatively massive amounts of use in preference over subgenres. Fourth, the indexes of ‘use warrant’ and ‘user warrant’ as constructed for this study very strongly correlate, with Spearman’s rank correlation coefficients of over 0.999 for both non-LCGFT and LCGFT orders.

The thematic analysis of headings in the WorldCat records for works tagged with the non-LCGFT terms did not reveal any clear substitute terms that were consistently used instead of these terms. For the most part, however, the headings indicated that the tags used in LibraryThing did indeed represent the genres described in Wikipedia and, in the case of pulp, as used on Ebay. There was one very noticeable exception, however, namely ‘northwestern’, which appeared to be used to describe fiction of a quite different ilk from the ‘western’ set further north in the American continent as per Wikipedia. In some cases, for instance, the library headings pointed to settings in Russia. The wide range of headings found in the records for works tagged ‘inspirational’ and ‘weird’ also pointed to a ‘slippage’ in meaning from that described in Wikipedia, although in these cases it is less clear to what extent the difference is absolute or one of degree. Many of the headings for ‘weird’ did relate to ‘the other’ and alienation, while many of the headings for ‘inspirational’ could be considered emotive.

The headings for the other nine corresponding genres are summarized as follows: chick lit works attracted relatively large numbers of headings, representing many different subjects and various categories of women; pulp fiction also had headings covering a wide variety of subjects, including fictitious characters, and sex and crime related topics; hardboiled fiction mostly had specific headings pertaining to characters, plot and place; gaslamp fiction had a large number of headings for classes of persons and some for broader genres, including science fiction; grimdark fiction also had headings for broader genres, as well as for a number of subjects with a military focus; dying Earth fiction had fewer subject headings, but headings for both fantasy and science fiction;
splatterpunk fiction also had fewer headings, tending to cover subjects related to violence; biopunk had a wide range of subject headings, including some with a scientific slant; and dieselpunk fiction had headings for a range of ‘dark’ subjects, but lacking any obvious common thread.

Sampling the headings, and the works described by these headings, it is clear that many of the non-LCGFT tags represent fiction that may be about various subjects, but that also, critically, invokes certain feelings, and, overall, more so than does the fiction tagged with the LCGFT genres, many of which indicate particular settings (e.g. prison fiction) or particular themes (e.g. social problem fiction), more than particular reader experiences. Even ‘chick lit’ and ‘pulp’, although their more literal meanings do not represent feelings, have widely known connotations that include affect. Indeed, Fenkel (2019: 183) has argued that chick lit forms part of ‘a type of umbrella genre’ that she labels as ‘pleasure’, ‘constituted as an ephemeral archive that is translated into popular fiction when it is read as a history of feeling in public cultures.’ It seems that pulp fiction might also form part of such an umbrella genre. In summary, the sampled non-LCGFT genres can be regarded as, on the average, more affective than the sampled LCGFT genres.

Although the method used in this study does not account for synonyms used by LibraryThing taggers for the terms under analysis and so strictly compares the use and user warrant of terms rather than concepts, it is nevertheless able to highlight, at the very least, deficiencies in language that could be of interest to vocabulary builders, and it could be extended to compare such deficiencies across multiple schemes. It can also be argued that most terms used in most schemes would be the most used or only commonly used term for their concept, and as such the index provides, even at the conceptual level, a reasonable ‘rule of thumb’ for the purposes of broad comparison.

6.0 Conclusion

The findings reported above strongly suggest that although libraries are describing fiction using some of the genres heavily used by the general population of fiction readers, they are also missing a considerable number of others. The greater amount of user and use warrant for the Wikipedia and Ebay genres than for the LCGFT genres, on the average, would also suggest that the impact of library classification on everyday genre classification (at least with respect to fiction) is weak relative to other classifications, such as those of experts via Wikipedia and commercial classifications. Therefore, while it may be that both ‘professional’ and commercial classifications strongly impact art classifications as a whole, some of these classifications are likely to be considerably more impactful than others. The more heavily used genres in this study tend to connote a strong affective element. The importance of affect for fiction searching has been noted by authors such as Stover (2009) and Mikkonen and Vakkari (2016). It is very possible that this element has, in fact, worked against the inclusion of some genres in library vocabularies. In some cases, the connotations may be seen as derogatory, but, beyond this, recognition of affect may have not sat well in the modernist, essentialist paradigm of which the Anglo-American library cataloguing has long been a part, with cataloguers viewing themselves as providers of access to information objects, rather than postmodern facilitators of access to materials that are engaged with perhaps primarily for emotional and subjective reasons. The modernist paradigm is therefore especially imperfect,
it would seem, for the pro-vision of access to works of the imagination, such as fiction. Rightly, the Library of Congress has separated out LCGFT from LCSH, as genres go well beyond subject. However, this study demonstrates that the genre list may need further work, underpinned by greater recognition of the importance of affect in the consumption of creative works, including the reading of fiction.

Acknowledgement
The author wishes to thank Anya Smeaton for her assistance in compiling the data for this paper.

References


BARTOC FAST
A Federated Asynchronous Search Tool for Remote Vocabulary Access

Abstract:
In this paper we introduce BARTOC FAST, a federated asynchronous search tool for remote vocabulary access. We first motivate the need for BARTOC FAST by exposing the limitations of the local BARTOC.org Skosmos instance. We then discuss the advantages of BARTOC FAST – vast search space, low footprint, and modularity – and provide an overview over its implementation and some design challenges. We close by considering some anticipated use cases and plans for future development.

1.0 Limitations of BARTOC.org

The Basic Register of Thesauri, Ontologies & Classifications (BARTOC.org) is a full terminology registry for knowledge organization systems (KOS). We aim to collect as many controlled vocabularies as possible in one place, to describe them uniformly and to make them accessible. So far this is not the case, at least not with regard to the searchability of vocabulary content. BARTOC FAST attempts to supplement BARTOC.org with an application that allows to search for concepts and terms in vocabularies. The whole project stems from our conviction that KOS presented in BARTOC.org should comply with the FAIR principles for the use of controlled vocabularies.

BARTOC.org consists of two main modules: one contains the metadata of currently about 3,200 KOS, the other provides the members of the KOS, i.e. concepts and terms, as long as they are available in SKOS format. For the SKOS vocabulary service we use Skosmos, an open source web-based SKOS publishing tool developed by Suominen et al. (2015).

Skosmos allows users to browse SKOS vocabularies alphabetically or hierarchically and offers structured concept display. For BARTOC.org, the most important feature was the global search, which enabled searching for concepts across all hosted vocabularies. Consequently, vocabularies from other terminology registries had to be uploaded to our own RDF Triple Store. This caused serious problems after some time:

- Skosmos could no longer process the large number of 1,436 vocabularies (3,314,740 concepts, 11,492,219 terms, with the massive Getty vocabularies not included).
- Since our vocabularies were mostly clones of remotely hosted vocabularies, it was a considerable effort to keep them up to date by periodically checking for updates or new versions. Comparable portals such as Linked Open Vocabularies face similar problems and solve them through manual annual reviews and comments in a separate metadata field.

1 https://bartoc.org
3 https://lov.linkeddata.es/dataset/lov/vocabs/airs
• Not all vocabularies were available in full SKOS.

We had to realize that Skosmos is a great tool, but was no longer sufficient for our special purpose of a powerful search instrument for plenty of KOS. Or as Osma Suominen puts it in the Skosmos User Forum: “Skosmos works well up to around tens of thousands of concepts (...). With more than 100,000 concepts, it starts getting slow” (Suominen 2015).

One could add that with millions of concepts the global search function collapses. However, smaller Skosmos deployments containing unique vocabularies, run by institutions like UNESCO⁴, Food and Agriculture Organization (FAO) of the United Nations⁵, National Library of Finland⁶, University of Oslo Library⁷, Inist-CNRS⁸, etc. provide REST-style APIs and Linked Data access to the underlying data. Our task was to replace the global search of Skomos with a federated search method that was capable of querying multiple REST APIs and SPARQL endpoints simultaneously to make millions and millions of concepts from any number of terminology registries accessible with one tool.

2.0 BARTOC FAST

We have overcome the aforementioned limitations of BAROC.org by developing BARTOC FAST⁹. BARTOC FAST is a remote retrieval aid for thesaurus, ontology and classification concepts. The acronym FAST stands for Federated Asynchronous Search Tool and should not be confused with OCLC’s vocabulary of the same name, which resolves to Faceted Application of Subject Terminology¹⁰. At present the FAST federation contains more than 20 resources – including BARTOC.org Skosmos¹¹, Getty Vocabularies¹², the Integrated Authority File (GND) via lobid-gnd, as introduced by Steeg et al. (2019), and the Research Vocabularies Australia¹³, see https://bartoc-fast.ub.unibas.ch/bartocfast/about for a full list – thus comprising a vast search space.

BARTOC FAST offers three decisive advantages over BARTOC.org. First, Skosmos can take a back seat and exclusively serve as an instance for SKOS vocabularies that are not hosted anywhere else. Secondly, the data in BARTOC FAST is always up-to-date, as it comes directly from the APIs of the terminology registries. Conversely the footprint of BARTOC FAST as compared to BARTOC.org is massively reduced since the maintenance and support of the terminology registries is delegated to their providers. Thirdly, BARTOC FAST is a single access point that allows users to search not only for vocabularies but also in vocabularies. The new search interface of BARTOC.org will combine both functionalities (see Figure 1).

---

¹⁴ http://vocabularies.unesco.org/browser/
⁵ http://agrovoc.uniroma2.it/agrovoc/
⁶ https://finto.fi/
⁷ http://data.ub.uio.no/skosmos/nb/
⁸ https://www.loterre.fr/skosmos/
⁹ https://bartoc-fast.ub.unibas.ch/bartocfast/
¹⁰ https://www.oclc.org/research/themes/data-science/fast.html
¹¹ https://bartoc-skosmos.unibas.ch/
¹² https://www.getty.edu/research/tools/vocabularies/
¹³ https://vocabs.ands.org.au/
Figure 1: Prototype of the new search interface for BARTOC.org including BARTOC FAST.

2.1 Implementation

BARTOC FAST is implemented in Python 3.7 and runs on the servers of the Basel University Library. The frontend and backend will be discussed in turn. The BARTOC FAST frontend uses the Django framework\(^\text{14}\) and comes with three distinct views: basic, advanced and API. Each view corresponds the expectations of a specific type of user:

1. The basic view mirrors the user interface of a discovery tool such as Primo\(^\text{15}\). All parameters (except the search input of course) are fixed and set to values that yield usable results in most cases. The user simply types in a search word and receives a list of results. This list is both sortable and searchable (see Figure 2) and hence allows for additional refinement.

2. The advanced view is similar to the basic view but allows for a customization of all parameters (see Figure 3). These include the maximum search time, the choice of queried resources, and the option of keeping duplicates. Note that some desirable facets, filters and modifiers are not yet implemented but scheduled for the next development cycle.

The BARTOC FAST API is a HTTP-based RESTful API, soon to be compliant with the Reconciliation Service API\(^\text{16}\) returning results as JSKOS, a data format for KOS by Voß (2019), or generic JSON-LD. The view itself is identical to the advanced view.

\(^{14}\) https://www.djangoproject.com/
\(^{15}\) https://www.exlibrisgroup.com/products/primo-discovery-service/
\(^{16}\) https://github.com/OpenRefine/OpenRefine/wiki/Reconciliation-Service-API/
Shifting the focus towards the BARTOC FAST backend now, since BARTOC FAST is a remote service, it needs to process many distinct APIs. For this reason, resource modelling and query resolving in BARTOC FAST is handled by a GraphQL\textsuperscript{17} schema which makes use of the Graphene module\textsuperscript{18}. Simply put, GraphQL provides a (meta-)query language over all the resources in the BARTOC FAST federation (see Figure 4). Resource modelling is modular and on an API basis. This means that new instances of already modelled APIs (such as Skosmos and SPARQL) can easily be added to the BARTOC FAST federation.

\textsuperscript{17}https://graphql.github.io/graphql-spec/June2018/
\textsuperscript{18}https://graphene-python.org/
Figure 3: The BARTOC FAST advanced view enables parameter customization.

Figure 4: A BARTOC FAST search request in the GraphQL query language.

BARTOC FAST resolves federated queries by translating the user’s search input into an API call for each resource; the list of aggregated results is then returned. More precisely, query resolution takes three steps for each resource: asynchronously fetching data by means of an API call to the resource, normalizing this data, and purging duplicates. Every result in the list of aggregated results has three mandatory data fields: a URI, a nonempty set of labels, and a source:

1. The URI is used to tell apart results within and across resources. Given the results of a single resource, all results with the same URI are merged into a single result and their labels (respectively the contents of their labels) are aggregated. For the results of multiple resources, redundant results as identified per URI are purged.

2. BARTOC FAST results use four SKOS labels\(^{19}\) as base, namely skos:prefLabel, skos:altLabel, skos:hiddenLabel and skos:definition. Since not all resources in BARTOC FAST employ SKOS, the semantic equivalents of these

\(^{19}\) https://www.w3.org/2009/08/skos-reference/skos.html
labels are provided via a mapping (e.g., skos:prefLabel is equivalent to rdfs:label\textsuperscript{20} is equivalent to gnd:preferredName\textsuperscript{21}). Note that the descriptor takes no special priority.

3. The source is the queried resource.

2.2 Challenges

In this section we discuss two design challenges that we encountered in the development of BARTOC FAST. An initial challenge concerns the construction of federated queries. Generally speaking, a federated query processes the input and passes it to each resource in the federation in a format that is accepted by the resource. So when a BARTOC FAST search request is carried out, the user’s search word is transformed into an API call for each resource in the BARTOC FAST federation. Since each API is represented as a model, this task is reduced to transforming the input string into a search request for each model. However, different models allow for different kinds of search requests. To give two examples: not all models support Boolean operators, and different labels are given varying importance in different models. Our solution to this challenge is twofold. First, instead of trying to construct formally equivalent search requests across all models, we opted for the pragmatic approach of constructing search requests with similar outputs and behaviors. Secondly, users can toggle a view of the exact API calls triggered by their string input on the BARTOC FAST results page. This whiteboxing awards users more control since unwanted API calls can simply be turned off. In the future we plan to include options for customizing federated queries (e.g., exact match over label X).

A second challenge concerns the varying sizes of the resources within the BARTOC FAST federation. Unsurprisingly, big registries such as the Integrated Authority File (GND) tend to outcrowd smaller registries with respect to the number of results for a search request. However, the number of results is not an indicator of quality. Generally speaking, there are at least two solutions available to this problem:

1. Expand or shrink the BARTOC FAST federation according to need. We have already implemented this solution by allowing the user to manually (de)select queried resources in the advanced view as discussed above.

2. Rank the results. Some simple variants for ranking include giving preference to label X over label Y, or giving preference to results with less empty labels. In addition to reducing noise, this solution has an additional advantage: if a resource is still dominant for ranked results, it is a marker of the resource’s quality rather than a problem. The downside of this solution is that the neutrality of the results is no longer guaranteed. The role of BARTOC FAST would shift from an aggregator to a curator. For this reason, if we decide to implement this solution, it will be strictly opt-in.

3.0 Outlook

At its current stage, BARTOC FAST is still a prototype with limited service, like a Google search kind of tool for KOS content. In this preliminary stage of development,
it does not fully benefit from all SKOS labels with the REST API queries or RDF data with the SPARQL queries, which allow different query types, plus the filters, aggregates, modifiers, and operators. Such extended functionalities will be part of future development steps.

However, with BARTOC FAST in production, we pay particular attention to possible use cases now. BARTOC FAST provides a valuable vocabulary resource for KOS mapping (e.g. via Cocoda mapping tool by Balakrishnan et al. (2018)) and automated subject indexing (e.g. via Annif by Suominen (2019)) of multidisciplinary digital repositories like Zenodo22 and similar, that don’t come with a controlled vocabulary natively. We also intend to further grow the BARTOC FAST federation by adding instances of already modelled APIs or by modelling APIs that promise to add value in scope or depth. If you know of a resource that should be added to BARTOC FAST, please do not hesitate to contact us! Finally, we plan to provide full access to the BARTOC FAST source code under a permissive license as soon as some security issues have been addressed.

References


22 https://zenodo.org/
Visibility, Identity, and Personal Expression
Qualitative Case Studies of Social Tagging on MetaFilter

Abstract:
Social tagging is often studied quantitatively and through a lens of tag typologies and terminological representation. This qualitative study examines three cases of social tagging on MetaFilter, a community weblog with extensive use of and discussion about tagging practices and the significance of tags. The cases illuminate the intersection of social tagging and the cultural and social themes of visibility, identity, and personal expression. We close with a reflection on how community affects tagging practices.

1.0 Introduction
Many websites such as Stack Overflow, GitHub, Goodreads, and LinkedIn allow users to label and organize information by using social tagging. The resulting folksonomies have most often been studied quantitatively, beginning with early studies that measured the distribution and types of tags on Web 2.0 folksonomies such as del.icio.us and Flickr (Kipp and Campbell 2006; Munk and Mørk 2007). While these and many other quantitative studies describe tagging behaviors and outcomes, they do not deeply explore how social tagging intersects with social and cultural themes such as visibility, identity, and personal expression in online communities.

We do have, however, a rich history of studying how other modes of subject indexing and classification intersect with these social and cultural themes. Bowker and Star (1999) describe how classification systems reflect and shape larger social systems and dynamics. Olson (2002) explores representation and identity in indexing and argues that the rigid vocabularies employed by early indexers like Dewey marginalize and exclude those who are different from the indexer. Tennis (2002) introduces the concept of subject ontogeny and traces how social and cultural changes affect our understanding of knowledge and classification choices over time. Furner (2007) applies critical race theory to study how ineffectively indexing languages represent racial minorities. Duarte and Belarde-Lewis (2015) examine cataloging and classification through the lens of colonialism and discuss the potential of Indigenous community-based approaches to representation and identity in information systems.

While these critical and social perspectives represent a major thread of knowledge organization research, only a small body of social tagging research has followed this path. The lack of attention to the social and community aspects of social tagging can be attributed, in part, to early social tagging platforms being weakly linked to online communities. For example, del.icio.us users with similar tagging histories could connect with each other and existing communities could establish canonical tags to share information with each other on services like del.icio.us (Tonkin et al. 2008); however, these social tagging activities were not strongly linked or integrated with other community activities. As social tagging has become a more integrated feature of online communities, a body of research has emerged that has, instead of treating the existence of social tagging as a social phenomenon in and of itself, studied tagging as an activity within these online communities that informs and is informed by cultural and social
factors. Much of this research has focused on terminological representation. For example, Adler (2009) compares the LibraryThing folksonomy with traditional indexes, focusing on representation and vocabulary in describing books with transgender themes. Adler argues that “[T]he greatest power of folksonomies, especially when set against controlled vocabularies like the Library of Congress Subject Headings, lies in their capacity to empower user communities to name their own resources in their own terms.” Bates and Rowley (2011) also consider visibility, identity, and personal expression in the LibraryThing folksonomy and find that minority voices can be well represented and that their identities can be accurately portrayed through concerted community effort. For example, they find that LibraryThing’s social tagging approach “offers benefits over LCSH... in the discoverability and representation of LGBTQ resources.” Coqc (2015) observes a community of Twitter users who tweet in Sami and use Sami hashtags specifically to raise the visibility of the endangered language. These hashtags function very much like social tags, and they allow these Twitter users to increase the visibility of an underrepresented culture and language. Cocq finds, however, that these tweets reach a limited audience outside of the small community producing them. Bullard (2016) captures user discussions about choosing preferred terms for social tagging in a fan fiction community. These discussions balance term popularity, utility for information retrieval, and potential harm to non-dominant user groups. Bullard’s approach allows us to see not just the terms that are ultimately used in the folksonomy, but community attitudes and values that inform the choices.

This study, taking particular inspiration from Bullard, aims to build on the broad tradition of knowledge organization research that considers the intersection of subject indexing with social and cultural factors. We also aim to add to the small body of this type of research that is specific to social tagging. The cases presented in this study consider, qualitatively, the themes of visibility, identity, and personal expression in social tagging that manifest in the community blog website MetaFilter and its subdomain for discussions about the site, MetaTalk. By analyzing these cases, we aim to address the following research questions: 1) How do social taggers use tags to increase representation and visibility of underrepresented voices in an online community? 2) How do social taggers use identification tags that respect the personal, social, and cultural identities of authors and content creators? 3) How do social taggers balance the utility of descriptive tags with the personal and social value of using tags for humor, commentary, and personal expression?

2.0 Method

This study considers three cases of MetaFilter community members discussing how to use social tagging effectively and appropriately in their online community. MetaFilter is a community blog that has run continuously for over 20 years, and is still active. MetaFilter has thousands of community members, paid staff and moderators, and numerous subdomains including a question-and-answer site called Ask MetaFilter. MetaFilter has global membership, but the language of the site is English. The United States and the United Kingdom are disproportionately well represented in the community.

Since 2005, authors of posts on MetaFilter and its subdomains have been able to tag or label their posts with as many tags as they like, using any vocabulary that they
like. Although moderators and members who are closely connected with the original authors can modify the tags for a post, MetaFilter is a narrow taxonomy, meaning that the post author provides the tags for their own posts and other community members do not contribute additional tags. The tags that post authors choose are particularly important in this folksonomy, then, because the community cannot establish different descriptions or identifiers through their own tags and because these tags are displayed prominently next to the content and are used for search and navigation throughout the site.

The MetaFilter community has discussed social tagging practices extensively, with 370 separate discussion threads and approximately 13,000 comments about tagging posted to the MetaTalk subdomain. These discussions are publically available and readily findable because they themselves are tagged with tags such as tags, tag, tagging, folksonomy, and labels. We collected all discussions with these tags, removed discussions about HTML tagging syntax, and began coding discussions by types of tags (Golder and Huberman 2006), motivations for tags (Gupta et al. 2011), types of initial post (feature requests, how-to questions, etc.), and topic of discussion (tagging syntax, culturally appropriate tagging, etc.). Key themes emerged during coding, and we identified particularly interesting discussions about visibility, identity, and personal expression that we surface in the following cases.

3.0 Cases

In this study, we consider three cases from the large set of tagging discussions on MetaTalk to illustrate and explore emergent themes from ongoing coding and analysis work. Two of the selected cases are contained completely or predominately in single discussion threads. The other case spans multiple discussion threads that are thematically connected. These cases were chosen because they illuminate key questions and themes about social tagging and folksonomies, they explore questions that are difficult to answer with quantitative analysis, and they feature rich and thoughtful discussion about how tagging intersects with concepts of visibility, identity, and personal expression.

3.1 #JulyByWomen and diverse global voices

The MetaFilter community regularly reflects on itself as an online community, aiming to create a welcoming space with open discussion and diverse voices. The #JulyByWomen campaign arose from discussions about women being underrepresented in the MetaFilter community, particularly as authors of posts (viggorlijah 2014a). The campaign encouraged women to author more posts during the month of July 2014 and to use the tag JulyByWomen on those posts to raise the visibility of women on the site. The campaign diverged from MetaFilter's typical use of tags because the JulyByWomen tags reflected the identity of the author instead of describing the content of the post. Despite—or because of—this divergent use of tags, the campaign was considered a major success. #JulyByWomen increased visibility of and participation by women (viggorlijah 2014b), and the use of a consistent tag also provided an indexing benefit, as all of these posts were automatically collected in a single place (“Posts tagged with julybywomen” n.d.).
The success of #JulyByWomen led to further discussion about using tags to increase visibility for underrepresented voices and topics on MetaFilter. The case that we consider in this study occurred in a MetaTalk thread titled “#GlobalVoices/#NonWestNov/#GlobalSouthSept,” which was posted on August 5, 2014, directly after the conclusion of the #JulyByWomen campaign and in consultation with, viggorlijah, the MetaFilter member behind #JulyByWomen (divabat 2014). The discussion ran through August 11, 2014, with 240 total comments made by 91 different users. The post author, who identified themselves in the discussion thread as an “Asian international student currently floating between countries,” aimed to introduce a new tag to promote posts by community members “that are outside the White Western norm, especially those outside the US” and “posts about people, places, and so on that take place outside the West” (divabat 2014). The discussants readily agreed with the premise that non-Western voices and topics were underrepresented on MetaFilter and with the goal of increasing their visibility. Community members also supported using a tag as an organizing principle for achieving these goals. However, the proposal received significant constructive criticism, and this criticism can inform our understanding of how tags and identity interact.

One criticism focused on a key difference in focus between #JulyByWomen and the proposed campaign for global voices. #JulyByWomen focused specifically on increasing visibility of women as authors of MetaFilter posts, while the proposed campaign aimed to increase visibility for both underrepresented post authors and underrepresented topics. This lack of focus caused confusion about the meaning of the proposed tag. Should Western community members use the tag to post about non-Western topics? Should non-Western community members use the tag when posting about general topics? The discussion did not reach consensus on how to support both proposed goals with a single tag. Additionally, the site founder, mathowie, and other community members expressed concern that encouraging specific subjects would make the campaign less successful than #JulyByWomen. The criticisms about focus suggest that successful tags, especially tags that communicate complex information like identity and culture, should not be overloaded with multiple meanings.

Another critique of the proposal was the lack of clarity in the proposed tags. Clarity is important because, while affording individuals freedom to choose their own vocabulary is considered a feature of social tagging in general, community members need a shared meaning for a tag to use it successfully in the context of a coordinated campaign. MetaFilter community members expressed confusion about the geographical and cultural boundaries of proposed tags like GlobalVoices, GlobalSouthSept, and NonWestNov. Some were unfamiliar with the term “Global South,” some were unsure of the boundaries of “The West,” some wondered why the United Kingdom might be lumped together with the United States, and some wondered whether non-white topics from Western countries fit the campaign. In contrast, #JulyByWomen had clearer boundaries. Despite gender being complex and non-binary, the concept of “woman” created less confusion than the cultural and geographical constructions discussed in this case.

Clarity is not the only rhetorical aspect of tag construction that matters for visibility. Positive framing is critical to making underrepresented groups visible, as negative framing can marginalize or “other” these groups. For example, JulyByWomen
was a positively framed and inclusive tag, while *JulyByNotMen* would have been a negatively framed tag and emphasized who was excluded. MetaFilter community members proposed a variety of tags for the global voices campaign. Those tags that used negative framing, like *NonWestNov*, *RestOfTheWorld*, and *BeyondUSA* were criticized for “othering” the very groups and topics that they aimed to promote. In contrast, positively or neutrally framed tags like *PostMoarGlobal*, *SeptemberForTheWorld*, and *GlobalFilter* were more considered more inclusive and more clear about boundaries.

Finally, this case shows that visibility has a temporal component. #JulyByWomen succeeded in part because it had clear start and end dates. One proposal for the global voices project suggested that there be an ongoing effort to raise visibility of these voices and topics. This idea was roundly rejected as a form of segregation by the post author and not revisited by other community members.

“Having it just be a 'use this tag when you're talking about stuff outside the US' kind of defeats the purpose of this project, which is to make a concentrated effort to highlight and showcase material from all around the world. It can continue past the month, but right now just having it as a general-purpose tag feels like it's siloing off those posts even more” (divabat 2014).

Ultimately, the global voices campaign did not achieve the success of #JulyByWomen, with no officially adopted tag and only 17 posts tagged with the most popular proposed tag, *PostMoarGlobal* (“Posts tagged with postmoarglobal” n.d.).

### 3.2 Post tags and deadnames

In this case the tag is an author's name, Daniel M. Lavery (Wikipedia contributors 2020). This author has published work under the names Mallory Oldberg, Daniel Mallory Oldberg, and Daniel M. Lavery. The question put to the discussion thread is two-fold.

“I posted Daniel Mallory Ortberg's latest instalment of his ongoing serial fic. When tagging the post, I was conflicted about tagging it with his deadname (which is "Mallory Ortberg"). Two questions: 1) on a technical level, will tagging with his full name link up with older posts tagged with his deadname, since the latter is included in the former? He's the same author, and that continuity of work seems valuable. 2) on a trans-etiquette level, is this shitty and equivalent to deadnaming? (I'm more interested in hearing what trans folks have to say on this one.) I ended up tagging the post with his name and his deadname but I'm questioning that” (sixswitch 2019).

The first question shows us a mental model on how tags work in MetaFilter. The second question is framed as an etiquette question, but given the responses in the thread it is not limited to that. For this tagging community it becomes a contested conceptual move. The term deadname is called into question as inappropriate. So we have two levels of terminology work that surface. One in the tag and its content and the other what we call part of the content of that tag. Both of these rely on community input to reconcile.

The first question is immediately answered. It is a technical question of the mechanics of the tagging system of the site. Further, action is taken by a moderator, Eyebrows McGee, to add the tag *danielmalloryoldberg* to all posts that had a combination of those names or a subset of those names.

The second question, while framed as an etiquette question, quickly turns into a discussion of the use of *deadname* in the post. The community rejets this terminology for a name used by a person before they transition to another gender. A decision is taken to flag tags like this as “flag with note” (Lobstermitten 2019). What follows are various discussions of the dynamics of interacting with trans people both in the community of MetaTalk and through tagging.
This etiquette question goes right to the heart of identity and its relationship to tagging. It required the community to engage with sensitive terminology and negotiate how identity changes and does not change in the context of gender transitioning. This engagement and negotiation required intellectual work to understand the role of tags in identity representation. The community did not demand change other than that taken by the moderator. Further, there was no resolution on best practices that the community adopted in this context. It remains to be seen whether best practices will surface around this particular issue at this time.

3.3 Personal expression in tagging

Tags on MetaFilter are valued primarily for their utility, as evidenced by recurring questions about tagging best practices and requests for search and browse features to increase tag utility. For example, MetaFilter community members ask about how to ensure that tags accurately describe the subject matter, how to make tags sufficiently specific, and how to format tags syntactically and orthographically to support retrieval in search and browse modalities (Going to Maine 2016). MetaFilter community members, many of whom are academics and librarians, answer knowledgeably about indexing best practices and the technical details of MetaFilter's tagging and retrieval ecosystem.

Despite the primary focus on utility, tagging is not seen as exclusively utilitarian by the MetaFilter community. Because tags are assigned by post authors, they are considered part of that post author's personal expression. Exercising the freedom that this perspective affords, some community members use tags to joke about and comment on the topics of their posts. These personal expression tags generally do not support retrieval and are rarely descriptive in an traditional indexing sense, but they do allow community members to express their personality and viewpoints and, when used appropriately, can build a sense of community. The question, then, becomes: When and how are personal expression tags beneficial, when and how are they inimical, and how does a community balance these potential benefits and harms?

Multiple MetaTalk discussions recognize the benefit of the creative use of tags. For example, one post asked community members what their favorite tags are and received numerous spirited responses about obscure and humorous tags (Fizz 2017). The humorous use of tags is evident throughout MetaFilter, such as one post with 75 tags, all synonyms for nonsense (Just this guy, y'know 2017). These tags entertain, provide commentary, and establish the personality of the community.

However, some community members have asked whether humorous tags are harmful. For example, one member asked whether an extremely long joke tag should “be frowned upon” (i love cheese 2006). The consensus response was that such tags are not harmful to the folksonomy and have value outside of indexing: “I don't think it affects the usefulness of tags, as long as more accurate tags are included. Plus it makes me smile” (cali 2006). This sentiment is repeated in a discussion about a more controversial tag, batshitinsane (UKnowForKids 2005).

“Silly tags are only noise when that's the only tag you use for your post. If there's four other decent tags included that people might actually use for a search, then why does it matter if there is one that no one will ever use as a search term?” (23skidoo 2005).

However, the tag batshitinsane is problematic for reasons other than being noise in the folksonomy—it is derogatory, vulgar, and persistently popular as a form of humor. As
such, it has been the topic of multiple discussion threads. One such post asks if the term is offensive (CCBC 2010). MetaFilter moderators determined that the tag should not be outright banned, but encouraged community members to use it responsibly. They considered gratuitous use of the tag in posts about mental health and use of the tag to editorialize and stifle discussion to be irresponsible uses, showing that in this case community norms and not utility determined the appropriateness of a personal expression tag.

Another form of personal expression through tagging is not about humor, positioning, or performance. It is about identity and visibility. Shortly after MetaFilter implemented tagging, one prolific user tagged all of their posts with their username to increase their visibility. This approach to tagging ran counter to the intent of tags being descriptive of content, but some community members considered it “harmless fun” (calwatch 2005). Site founder mathowie, however, considered this type of personal expression through tagging to be harmful.

“Tags were added as a way to categorize everything on the site under descriptive keywords. The poster's name doesn't really impart any info, and since you can already find every post by a username, I repeat that it's in effect already built-in and pointless to essentially state the same information twice. There's no need to make usernames into explicit tags. I'm sure quonsar just wanted to see his name in lights on the popular page, which he did, and now is gone. Whoop-de-doo” (mathowie 2005).

This case of tagging as personal expression was considered harmful for reasons of utility—it added confusing noise to the folksonomy—and for reasons of community norms—it did not build community through personal expression, instead benefiting just one user.

4.0 Findings and Conclusion

As is reflected in the literature, tagging is an activity that expands our notion of indexing (Golder and Huberman 2006; Tennis 2006; Munk and Mørk 2007). Tags are powerful symbols of community and are infused with the power to create inclusivity in the community. Therefore, care must be taken to understand how best to curate a tag collection to promote visibility, identity, and personal expression in the community. The power of tags, and the care that they require, are reflected in the extensive discussions on MetaTalk. We analyzed three cases in these discussions to study, qualitatively and deeply, how the MetaFilter community uses the power of social tagging to increase representation and visibility of underrepresented voices, to respect the identities of authors and cultural groups, and to balance the utility of descriptive tags with humor and personal expression.

We found that tags can increase visibility for underrepresented groups, provided that the tags have a clear focus and purpose, clear boundaries, positive and inclusive framing, and are part of a time-bound, concerted campaign. Especially in folksonomies that support unlimited tags for each resource, these visibility tags do not harm retrieval utility and can have a significant benefit of inclusivity, both practical and symbolic. However, the community must agree on a canonical tag for such efforts to succeed. The effort to promote voices outside of the Western and white voices that dominate the community failed because the community, despite significant care and effort, could not establish a clearly defined canonical term to use as a tag.

Similar to our first case, which showed that care and attention can increase the visibility of women's contributions on MetaFilter, our second case demonstrated that
substantive and careful effort can ensure that a social tagging community respects a trans author's wishes in naming. Community moderators implemented a technical solution to unite works authored under different names, and the community discussed with care both the intersection of social tagging and personal identity, and the ethics—not just etiquette—of deadnaming and the use of the term “deadname” itself. This discussion illuminates the symbolic and practical power of tags as signifiers of identity, and shows the importance of not tarnishing tags by implementing them without care and attention.

Finally, in the third case, we saw the tension present between utility and humor in tagging. The role of humor was both contested and celebrated in this context. The community viewed potentially superfluous tags as not harmful, provided descriptive tags were also provided, and beneficial to personal expression, humor, and a sense of community. Specific instances of humor were called out as inimical, however. The use of humorous tags to trivialize sensitive subjects, to stifle balanced discussion, or to promote one's self were considered inappropriate and harmful uses of the power of tags. It seems, then, that using tags for personal expression and humor are generally supported, provided their use does not undermine the core utility of tags or the social values of the community.

The deep, qualitative analysis of social tagging that is presented in this study represents a small step toward understanding social taggers in a more robust way, akin to how we understand the motivations and knowledge of professional indexers. Through the vibrant communication among members of the MetaFilter community, we can see the attitudes, motivations, knowledge, and values that shape social tagging in an online community. We recommend further study of modern social tagging sites to better understand how social tagging works as an integral feature of online communities.

References
https://metatalk.metafilter.com/9899/Batshitinsane-Tags#235619
https://metatalk.metafilter.com/8863/someone-likes-the-sound-of-his-own-name#193597
divabat. 2014. “#GlobalVoices / #NonWestNov / #GlobalSouthSept.” MetaTalk.


Gupta, Maniah, Rui Li, Zhijun Yin, and Jiawei Han. 2011. “An Overview of Social Tagging And Applications.” In *Social Network Data Analytics*, edited by Charu C Aggarwal. Boston, MA: Springer, 447-497.


viggorlijah. 2014b. “#JulyByWomen slightly passes goal.” MetaTalk.


Numerical Classification and Complexity
Developing a Classification of Classifications

Abstract:
The difference between monothetic and polythetic classification is well established in the literature (Sneath, 1962; Sokal & Sneath, 1963; Needham, 1975). Monothetic classification defines a group such that all members share specific common features, and that, at least in regard to their defining characteristics, any member of the group is substitutable for another. Polythetic groups, on the other hand, are “composed of organisms with the highest overall similarity, and this means that no single feature is either essential to group membership or is sufficient to make an organism a member of the group” (Sneath, 1962, p. 291). Numerical classifications, per Sokal and Sneath, are defined as a type of polytheticism. We argue that polythetic and numerical classification are not coterminous, and that all three classifications vary along an axis of complexity. Distinguishing characteristics of complexity include the number and nature of membership criteria, the internal structure of a classification, and the nature of consensus used in determination of a classification.

1.0 Introduction: Monotheticism, Polytheticism, and Complexity

The difference between monothetic and polythetic classification is well established in the literature (Sneath 1962; Sokal and Sneath 1963; Needham 1975). Monothetic classification defines a group such that all members share specific common features, and that, at least in regard to their defining characteristics, any member of the group is substitutable for another. Polythetic groups, on the other hand, are “composed of organisms with the highest overall similarity, and this means that no single feature is either essential to group membership or is sufficient to make an organism a member of the group” (Sneath 1962, 291). Monotheticism is commonly attributed to Aristotle (Topics; see, e.g., Hjørland 2017); polytheticism a creation of Wittgenstein (Philosophical Investigations). Lakoff also attributes polytheticism to the work done by Rosch in psychology, a body of work she called “prototype theory.”

In what ways can a classification be said to be complex? On the one hand, classification generally represents a reduction of the universe—a simplified model—by substituting a smaller number of kinds for a larger number of individuals. The substitution of types for tokens also foregrounds important or essential features—definitional characteristics in the parlance of monotheticists—so that not merely the number of things perceived is reduced, but the act of perception itself is concentrated on meaning-bearing features, with other features rendered to the background or discarded as inessential or non-salient. Finally, the identification of essential or salient features allows for cognitive economy. Categories (or the prototypes that reside at the center of polythetic groups) foreground a series of quick mental operations including recognition, memory, and making inferences regarding classified individuals or objects. For example, classifying something as metallic allows for expedited reasoning regarding the object’s look and feel, its hardness and weight, durability, and the possibility that it could be magnetic. Exceptional metals, such as mercury or sodium–potassium alloy (NaK), which are liquid at room temperature, are handled as mental
exceptions. Such categories, including flightless birds, while exceptions, still represent simplifications of the world insomuch as a general category and its exceptions is a reduction in complexity from mentally listing all the subordinate species of birds, much less the extraordinary social and psychological burden of treating all birds as unspeciated individuals.

While a classification is by its nature is a simplification, it can be simple or complex in relation to another. Monothetical classification, as a generalization, can be said to be simpler than polythetic classification. Once we have consensus regarding membership criteria for a class, the process of assigning individual cases to the class can be relatively simple. For example, osteoporosis is defined as “A value of [bone mass density] 2.5 standard deviations or more below the young adult mean (T-score ≤ –2.5)” (WHO Scientific Group on the Prevention and Management of Osteoporosis 2003, 57), and further provides descriptions of standard diagnostic procedures and reference standards for generating T-scores. Clear definitional criteria provide simple binary membership rules. The internal organization of monothetic classes are also comparatively simple. Once an individual meets membership criteria, they are in the class, and all members of the class are substitutable for each other in reference to those definitional characteristics.

This relative simplicity of monothetic classifications isn’t to suggest that all such classifications are easy to construct. In fact, because the membership criteria are of high consequence—disallowing for partial membership at the margins of the class, for example—it can be difficult to specify those criteria so that the resulting class includes what we want to include, and excludes what we have habitually or tacitly excluded. Bennett (1980) and Gould (1981) discuss the construction of the class “zebra” which traditionally refers to several extant and extinct species of the genus Equus. Gould asks whether “they form a true evolutionary unit” (6) and finds that the class as popularly constructed may not meet the cladistic requirements derived from evolutionary theory. How do we proceed? Do we admit that we have irreconcilable differences on the construction of the class and arbitrarily choose one method for developing definitional characteristics, or do we gerrymander the class in such a way that it includes what we want to include and rig the rules of definition so we get the right result?

And just because we have definitional characteristics doesn’t mean they are easy to apply. The Miller test (United States Supreme Court 1973) provides the following definitional criteria for the class “obscene materials” (24):

a. whether “the average person, applying contemporary community standards,” would find that the work, taken as a whole, appeals to the prurient interest …;

b. whether the work depicts or describes, in a patently offensive way, sexual [or excretory] conduct specifically defined by the applicable state law; and

c. whether the work, taken as a whole, lacks serious literary, artistic, political or scientific value.

To be considered obscene, all three criteria must be met. The third criterion, known colloquially as the SLAPS test, is, like many legal definitions, notoriously complex in interpretation and application, requiring an assessment (inter alia) of literary value. The U.S. Internal Revenue Code of 1986 (United States 2018) is primarily a classification that assigns people into tax brackets—at 3,842 pages, it can only be characterized as a highly complex monothetic classification.

However, despite their potential complexities, monothetic classifications are generally simpler than polythetic classifications. Like a good theory, monothetic
membership criteria are designed to be parsimonious in nature and limited in number, and when well designed, are made with an effort to find clear and meaningful distinguishing markers of membership. Monothetic criteria are designed to yield clear outcomes—either in or out—and therefore the internal structure of a class is monotonic: in regards to membership criteria, each member of a class is substitutable for another.

The standard description of polythetic classification is given in Wittgenstein's notion of family resemblances which describes a nuanced yet relatively simple form of polythesticism based on a limited number of generally well-understood though perhaps hard-to-measure criteria. “We see a complicated network of similarities, overlapping and criss-crossing: similarities in the large and in the small … I can think of no better expression to characterize these similarities than ‘family resemblances’: for the various resemblances between members of a family—build, features, color of the eyes, gait, temperament, and so on …” (Wittgenstein 1953/2009, §66–67). The notion of family resemblances is perhaps understood as a discrete and small number of criteria whose similarities are best intuited rather than measured. Compared to classical monothetic groupings, polythetic classes are more nuanced and complex. Definitional criteria are potentially larger in number, with an appeal to a gestalt of formal patterns of overlapping criteria. Because not all definitional criteria are universally present and are understood best in relation to each other, the internal structure of a class generally includes membership gradation, with central and marginal members, and fuzzy membership boundaries. The distinction between monothetic and polythetic classifications along an axis of complexity is pretty straightforward: discrete definitional criteria, all of which must be present, versus a shifting array of definitional characteristics, incorporating nuances of similarity and difference.

2.0 The Construction of Numerical Classifications

Sokal and Sneath define numerical classification as “the grouping by numerical methods of taxonomic units based on their character states” (1973, xii). For Sokal and Sneath, numerical classification is a type or subset of polythesticism: it is polythesticism with empirical observations, concrete measurements, and statistical assessments of similarity and correlation—and an appeal toward scientific justification, or alternatively, scientific pretense. Montoya (in press) states that numerical classification is the grouping of entities (organisms, documents, data, etc.) using quantifiable measures of evaluation (such as traits, terms, values, etc.). Numerical taxonomy includes the processes of selecting representative entities, weighting entity values, clustering entities based on these values, relating entities into clusters, and including them into a system based on uniform theoretical commitments. Strictly speaking, numerical classification entails that membership criteria be expressed in numerical form, and the monothetic classification defining osteoporosis is precisely of that nature.

We argue that while monothetic, polythetic, and numerical classifications contain simple and complex examples, they are not coterminous, and they can be viewed as existing at different positions along an axis of complexity. The purpose of our paper is to lead to the development of a classification of classifications. One dimension of this classification is to consider the complexity of the similarities or resemblances that form the basis of a classification. Eventually we will discuss additional modalities of classification, including Hjørland (1998; 2017) who differentiates classifications by
their “methodology” of class formation, or the various kinds of theoretical commitments implicated by various forms of classification. *Queer classification* in particular is a useful concept for understanding the implications of those theoretical commitments, and emphasizes membership criteria that are semiotic (i.e., interpretive and socially constructed) in nature. However, in this paper we will present an initial assessment of complexity, in relation to monothetic, polythetic, and numerical classification, to determine whether complexity is a useful criterion of comparison for classifications generally. Finally, we also believe that numerical classification is perhaps the least understood and least theoretically integrated into the literature of knowledge organization, and so the bulk of our remaining comments will be aimed to address it primarily.

### 2.1 Assessing the Complexity of Numerical Classification

While we have defined numerical classification above as strictly being numerical in nature, it has a number of additional features as it is presented in its canonical form by Sokal and Sneath (1963). The features used in numerical classification—which is aimed at the identification of species and their evolutionary development—are not merely *numerical* but also *numerous*. For example, Bennett (1980, 273–274) identified 21 features in her analysis of zebras, including, for example:

1. Number of functional digits;
2. Degree of isolation of protocone and hypocone;
3. Presence and size of preorbital facial fossa;
4. Presence and degree of development of secondary infundibular fold on I2;
5. Size and position of inferior canines;

... 21. Presence of opisthotic dolichocephaly.

Common uses of numerical classification incorporate large numbers of observations which form the basis of similarity that can only be assessed statistically via complex computational techniques. The feature sets can be quite large, and generally a feature is included if it is hypothetically relevant to class formation. Because features have uncertain relationship to any class, some features may be non-salient, and even redundant, in the sense that they strongly co-vary with another feature, marking the possibility that both features might be the consequence of some unidentified latent feature. These all indicate complexity in a potential classification, and are in contrast with polythetic classifications where features may be difficult to explain but all have at least tacit relevance to the class in question.

### 2.2 Consensus

How consensus functions in numerical classifications depends, in part, upon *when* in the process consensus is being implemented as a rubric for structuration and decision-making. If we look to consensus classifications in the biological or biodiversity world, as we see in the Global Biodiversity Information Facility (GBIF), for example, consensus is often used as a mechanism to create “taxonomic backbones” upon which data points for various species are appended (2019). In this case, consensus is used as a mechanism to provide a generically-agreed upon taxonomy that can then serve as an organizational and access mechanism for species data that, at their points of origin, may or may not have been contextualized within taxonomies with the same philosophical or
methodological commitments. Consensus backbones essentially present a universal structure to avoid the inevitable conflicts between one taxonomic opinion and another. The same kind of universal approach is used in many bibliographic systems as well, though often not in automated ways—the Dewey Decimal Classification (DDC) system, for example, uses disciplinary subject partitions to organize documents. The result is that with the DDC, as in GBIF, some class decisions are counter to some pockets of expert opinion. The reorganization of the rosid family of angiosperms is a case in point (Green and Martin 2013). Due to the prevailing popularity and use of phylogenetic approaches, the DDC found enough warrant to change the schedule to reflect new scientific approaches. Yet despite the fact that phylogenetic approaches are now preferred and accepted, this same schedule must be used to organize documents even not subscribing to this schema. Consensus as an organizational approach defers to the majority knowing that the scientific world of opinion is plural. Given this reality, entities placed within consensus systems should be understood to have several distinct, and perhaps conflicting, identities: on the one hand, they have their position within the consensus taxonomy, juxtaposed with other entities within an environment that is ostensibly more global in nature; on the other, one also must understand that, at its point of origin, that same entity may or may not have been constructed or contextualized on the same ontological terms. Complexity increases if we cannot wrest these two identities apart from one another.

So, while the DDC is certainly consensus-based, a body of editors, as well as particular cited evidence, or warrant, can be pinpointed as the source of this change—and thus is also a source of bias in a classification’s construction. In the case of GBIF, however, and other automated synthetic systems, consensus decisions are not so easily visible, nor are the arbiters of this change identifiable. GBIF’s backbone taxonomy is “updated regularly through an automated process in which the Catalogue of Life acts as a starting point also providing the complete higher classification above families” (GBIF, 2019). It is precisely this “black box” of automation that makes numerical approaches especially complex, more difficult to understand. If we think of a more embedded automated system such as the Google search engine, we see this phenomenon clearly: there is no way to understand how algorithmic techniques are intervening to propose certain top-level classes for a seemingly-endless corpus of online documents. In these spaces we can ask, Why do these terms (or traits, or phenomena, or entities) mean more than any other? How are entities and traits valued in relationship to one another? And, What were the alternate possibilities by which these results could have been interpreted to an equally-valid state?

Google is also a prime example of how the question of when consensus intervenes becomes important. For Google, or any other dynamic searching or retrieval mechanism, classifications are dynamic, which means that the principles for construction, quantities being examined, and the resulting classes, differ each time we query a classification. For example, searching for a complex phrase in Google one day can provide a different set of results than another day. This is because the commitments for classification construction and the body of possible entities are changing. Humans also intervene in these algorithmic structures in ways that we cannot totally understand. Safiya Noble made this readily apparent in her work on race and algorithmic power (Noble 2018). In 2012, Noble published an article in Bitch magazine noting the
marginalization and racial classification of “black girls” and women on Google’s interface. “By August 2012,” Noble states, “Panda (an update to Google’s search algorithm) had been released, and pornography was no longer the first series of results for ‘black girls’; but other girls and women of color, such as Latinas and Asians, were still pornified” (2018, 4). In cases like these, automated consensus mechanisms, and the constant rate at which they are applied, confound our ability to understand them: even if—and that is a big if—we manage to understand the logic of classification at one moment, the results may have very little bearing a week later.

Returning to the notion of consensus, it is necessary to state an obvious fact: consensus is not, despite rhetoric to the opposite, equatable to universal agreement—at least not in the case of classification. Any one person can contest decisions made through automated means and, in fact, a critical approach to this work would support and popularize this approach. We can perhaps go so far as to say that consensus may be more-or-less equivalent to authority, so far as we, the users, acquiesce in some way to the fact that so-and-so system will be authoritative in one situation or another. GBIF, for example, is an authoritative source for data, but it certainly makes no claims about agreement within the scientific community about the taxonomic perspective it proliferates. Problems arise, again as Noble shows, when the authority of systems becomes authoritative without a sense of critical analysis. Google results should hold no authority on the question and the formation of our racial, ethnic, or cultural identities, and yet, this is precisely how they are being used whether purposefully or not. This is a critical point to understanding, and limiting, the impending wave of classificatory systems resulting from the application of artificial intelligence solutions to big data, for example, in “smart city” projects. The complexity of data and the sophistication and apparent neutrality of algorithms result in decisions that bear the authority of the only-partially understood classificatory regime but whose actual heuristics and resulting classificatory decisions make little sense, and which fail to provide justification or alternative possible constructions. Unquestioned algorithmic complexity is a dangerous social reality and, as such, it makes good sense to delineate what we do and do not understand about these systems of organization.

2.3 Issues of Complexity: What We Know and What We Know We Don’t Know

Classification and knowledge organization’s long history of scholarship gives a good grounding to understand some of the basic known factors about classification that we understand to be fluid, contended, and arbitrary. The space here is not sufficient to mention them all, but some basic issues can be identified as they relate to numerical classification. When thinking about the quantification of factors in numerical and algorithmic taxonomic methods, we know well that the values we apply to attributes or entities are, if well intended, arbitrary nonetheless. Let us take the example of a phenetic classification of pine trees—a classification based on formal physical characteristics. There are some quantities that we might find important: that they are evergreen, the texture and structure of their bark, dimensions and characteristics of cones, needle count and position, height, etc. That these qualities are used to classify a pine tree is arbitrary in that we could have identified any number of what might be considered non-essential characteristics: flexibility in the wind, utility as fire-wood, etc. Likewise, when we think of the numerical classification of documents, a system may use terms, co-term
prevalence, document source, authorship, keywords, etc. In documentary analysis this aboutness is essential to description but also evasive in terms of method identification.

And then there are factors for classification that may be difficult to identify and measure. It wasn’t until genetic testing could identify sequences of importance (the COI or COX1 “barcoding” gene, for example) that phylogenetic approaches opened the door for revolutionary taxonomic methods in the biodiversity sciences, for example. We could finally “measure” organism classes in a way that was “universal” and replicable. In the bibliographical world, the notion of relevance (Wilson 1973) has always been identified as central to information retrieval and selection, and yet truly quantifying relevance in a way that meets the searching criteria for infinite moments of need still evades us. Relevance is the primary goal of search engines given that searches are explicitly intended to satisfy some situational need. And so we know that there are some qualities that are obvious, some quantities that are difficult to define, and in both cases, what we choose to seek out is wholly arbitrary based on our assumptions about the world, our ontological commitments, and our contextual purposes for organizing.

On top of the qualities we use to class entities, we must also insert our own hermeneutic skills to interpret their meanings as they relate to one another. Relationships are neither given, nor obvious, and will always depend on the context in which they should function. “To specify a relationship, we may first designate all the parties bound by the relationship (hereafter referred to as the participants in the relationship) and then specify the nature of any relationship that binds them together” (Green 2008). This means that relationships made by a scientist using phylogenetic methods, for example, will be based on specific and arbitrary quantities and articulated in equally arbitrary thresholds for a given set of taxa (even if the decisions are evidence based and properly “scientific”). But these constructed relationships are not natural relationships: they are imposed interpretive frames. In the end, why we make any given decisions can be based on clear guidelines or can be based on tacit or unconscious factors. In phylogenetics, clear mechanisms to distinguish one distinct species from another can be identified, along with the thresholds used to assess taxa. When Francis Galton was using composite photography to classify types of criminals in the last quarter of the 19th-century, however, it is clear that racial factors were taken into consideration. The history of the classification of race and humans is riddled with these conscious and unconscious biases (for example, see Smith 2015).

Some unconscious biases impact classifications that are less easily identifiable. Ontological commitments, for example, are sometimes difficult to archaeologically unearth in certain biological taxonomies without the producer on hand to explain certain decisions. Numerical classifications add a layer of complexity onto this that is significantly more complicated: the fact that statistical models are both mathematically complex and difficult to reverse engineer to understand class partitions at any given point in time. Once again, looking at Noble’s work (2018), an essential problem with automated organization is that it becomes very difficult to identify the location(s) of error when assessing a given set of results. Error in this sense is multi-valenced; it is locational (as in, there is ostensibly a code location and directive to locate), temporal (when, in fact, did this decision occur in a long scale of decision locations?), and multivariable (what variables or quantities were being referenced at that particular point and time?).
3.0 Function

Our paper has attempted to appraise three kinds of classifications in terms of their complexity. Monothetic, polythetic and numerical classifications are judged to be increasingly complex, by virtue of the number and nature of their definitional characteristics, their internal structure, the nature of consensus in their formation, and the comprehensibility of their resulting classes.

However, as we have also noted, there are simple and complex examples within each type of classification. While we believe that monothetic classifications are generally simpler that their polythetic and numerical counterparts, such a conclusion may be shaped in the way we have generally understood each type of classification. By choosing Sokal and Sneath, for example, to represent the canonical form of numerical classification, we may have unwittingly opted into a more complex version of that classification. We would not typically use the taxonomy of species as the primary and nearly exclusive use of any one type of classification.

Additionally, we are certain that we have not located all the various modalities of classificatory complexity. Definitional characteristics and the nature of consensus have figured prominently in our previous work on classifications, but there are certainly other salient facets to complexity that we have not yet explored. For example, we know that the automated and numerical classifications that predominate in common web-based applications are also closed and proprietary (e.g., Pasquale 2015), making them inherently more complicated to assess. This paper should be viewed as an initial foray into assessing the complexity of classification.

Additionally, we have been operating under an assumption that the three kinds of classifications presented here are substantially different in kind. There may yet be a theory that, for example, presents them each, in sequence, as a generalization of a previous model. That is, perhaps, that monothetic classification is a more specific kind of polythetic classification, with more precise definitional characteristics. That a given classification might be more complex may not, strictly speaking, mean that it’s actually different.

Finally, we need to consider that complexity ultimately may not be the right dimension for explaining differences amongst classifications. Another characterization, one that partially replicates the criteria used here or even one entirely novel, might be more effective at differentiating and understanding classifications. Complexity is simply our first, best stab at trying to understand classifications. Numerical classifications, in their automated and recent guises, are relatively entries in the history of classification, and accommodating them into our understanding of classifications generally is an unfolding process. Over time, the novelty of numerical classification—and the significance of its differences with antecedent models—may fade, and numerical classification may eventually be viewed as the same as its polythetic cousins. But right now, as automated approaches emerge in novel and not entirely welcome ways, with uncertain social and political consequences, we endeavor to understand what is new, and what is the same, with numerical and automated approaches to classification. Numerical classifications certainly feel inventive, riskier, and more complex. Their deployment, arriving as they do without complete user comprehension as to the nature of their operation, represents a new period in the history of classification, and their complexity masks uncertainty in the consequences of their use.
References


Deborah Lee – Department of Library and Information Science, City, University of London, United Kingdom
Lyn Robinson – Department of Library and Information Science, City, University of London, United Kingdom
David Bawden – Department of Library and Information Science, City, University of London, United Kingdom

Operatic Knowledge Organisation
An Exploration of the Domain and Bibliographic Interface in the Classification of Opera Subgenres

Abstract:
The classification of Western art music is notoriously complex, and the classification of opera subgenres provides distinct challenges. So, this paper considers the classification of opera subgenres from a knowledge organisation perspective. The paper starts with a short examination of key ideas from genre theory and musicological writings on genre, as well as the idea of opera subgenres. The categorisation of opera subgenres in the music domain is examined, utilising key music essays and sources. The large number of opera subgenres is identified, and the subgenres are explored using the framework of hierarchical, equivalence and associative relationships. The treatment of opera subgenres in eight bibliographic classifications is examined, where it is found to both reflect the disarray of the music domain and offer distinct discords. A model is proposed which considers the classification of opera subgenres on two planes, combining the web of relationships between subgenres (inter-subgenre plane) with the categorisation of the subgenre’s constituent attributes (categorisation plane).

1.0 Introduction
Opera is a significant part of the study and performance of Western art music. Yet, opera does not appear to have a systematic classificatory framework for its subgenres, suffering from unmanageable quantities and a lack of a standardised set of subgenres. While the overall facets of music have been studied in knowledge organisation (Elliker 1994; Lee 2017a), and the medium facet has received particular attention (Lee 2017b; Lee and Robinson 2018), the form/genre facet within Western art music has not been deeply analysed. So, this paper considers the classification of opera subgenres. It utilises knowledge organisation theories and concepts to explore the classification of opera within the music domain, and to compare this with the treatment of opera subgenres within bibliographic classification schemes. Hence, knowledge organisation will be employed to help understand and to disentangle the so-called chaotic nature of opera subgenres.

The paper starts with a short review of key ideas about music genres and opera subgenres, from the perspective of Western art music. Next, the classification of opera is explored from the perspective of the music domain. Some musicological sources are analysed to illuminate the quantity of opera subgenres and to consider the relationships between different subgenres. The treatment of opera subgenres in eight bibliographic classification schemes is considered, demonstrating interesting discords with the music domain. Finally, a model of classifying opera subgenres is presented, which explores how the categorisation of constituent parts of opera subgenres can interplay with relationships between subgenres.
2.0 Introducing genre-as-classification

Studies and analysis of the idea of genre have a long pedigree and many different domains are interested in conceptualising and utilising the idea of genre. Frow (2006, 10), working in the realm of critical theory and literature, defines genre as “… a set of conventional and highly organizing constraints of the production and interpretation of meaning”, showing how genre is concerned with structure and rules for description. Genre’s role as a way of distinguishing things and taxonomic function (Frow 2006) is discussed by genre theorists and those within the music domain (see for example, Holt (2007) writing about popular music). Although the role of genre goes far beyond a taxonomic device (Andersen 2015), this paper is concerned with the taxonomic idea of genre, sitting alongside other papers in knowledge organisation which discuss genre categorisation in artistic forms (for example, Rafferty 2010).

What is meant by a musical genre requires consideration. First, note that although the idea of a musical genre can have a wide range of meanings, in this paper genres are considered to be individual groups of works within the Western art music tradition, rather than the term “Western art music” being the genre. Second, the position of genre in the faceted classification of music can become blurred with form (for further discussion see Elliker 1994 and Lee 2017a). Third, it is useful to consider what elements make up a genre. Frow (2006), writing from a general genre theory perspective, considers a genre to be constituted by a number of aspects including formal features, thematic structure, physical setting, “situation” and more. Tereszkiewicz (2014) says that works of the same genre will have similarities in form, content and function. Noteworthy in these writings are the presence of form and function, themselves often considered facets of music. Frow (2006) also includes extrinsic qualities, showing how genre moves beyond the intrinsic qualities of a musical work. Music discourse also seeks to define a musical genre’s attributes. Dahlhaus (1987) is a useful source and suggests that genre consists of text, function, medium and form. (These aspects have been translated into standardised music classification terminology.) The idea of medium – who is playing or singing the work – is particularly important: Samson (2015), positions instrumentation as a defining feature of genre, while Dahlhaus (1987) defines genre as the expected connection between form and medium.

This paper is concerned in particular with subgenres. Subgenres are defined by the Oxford English Dictionary (“subgenre, n.” 2019) as “A subdivision of a genre of literature, music, film, etc.”. A brief perusal of music literature suggests that subgenre is a valid term for types of opera: for example, Carter (2014) and Senici (2014) use the term subgenre in essays about opera, and a key encyclopedia entry for opera (Brown et al. 2001) also uses the term when referring readers elsewhere. So, what does it mean to be a subgenre of opera? At its most literal, any specific type of opera will count. When and how a subgenre becomes a subgenre in its own right, and on whose authority, is an intriguing question. For example, opera seria is a significant subgenre of opera (McClymonds and Heartz 2001); however, these works were called dramma per music at the time they were written, with the term “opera seria” being adopted by those writing from a historical viewpoint (McClymonds and Heartz 2001). This illuminates how subgenre creation and categorisation of operas into subgenres can be enacted by those removed from the works’ creation, such as historians and theorists.
3.0 The classification of opera in the music domain

The first stage of this research involves analysing the music domain’s conception of the classification of opera subgenres; to do this, sources of information from the music domain are needed, which illustrate the classification of opera subgenres. Searching the literature does not reveal any standard knowledge organisation systems for subgenres of operas. Instead, the alphabetical list of 69 “see also” links in the Grove Music Online article for opera ((Brown et al. 2001) is a starting point – Grove Music Online (2020; henceforth abbreviated to its common name of Grove) is the seminal encyclopaedia and source for the study of music. Importantly, two musicologists writing generally about opera and genre, Campana (2012) and Carter (2014) utilise this list when reflecting upon opera subgenres, and indeed Campana (2012) refers to the list as a typology. The Grove typology (Brown et al. 2001) contains 69 terms, of which 66 are types of musical-dramatic works. However, the Grove typology presents some issues. First, Campana (2012, 221) refers to the typology as something produced “without any ambition of thoroughness”. This can be confirmed by the inclusion of three non-genre terms (verismo, libretto, Jesuits) and the exclusion of confirmed subgenres of opera. Second, the typology raises some questions about what is included in the boundary of opera, such as, whether works for dissemination through film are really subgenres of opera. Third, the typology is an alphabetical list of generic labels, so further sources are needed to contemplate the structure of opera subgenres.

So, supplementary sources will also be used. Campana (2012) also refers to the Wikipedia table of subgenres of operas. This Wikipedia table (“List of opera genres” 2019) is a useful resource: some of its subgenres are not found in the Grove typology and some entries include descriptions of their relationships with other terms in the table. It is also pertinent to supplement these list-like KOSs with ideas about classification not contained within an actual KOS. Two musicological essays which (briefly) discuss the classification of opera subgenres will be used: an essay on genre and poetics by Campana (2012) and an essay questioning the nature of opera by Carter (2014). Finally, Grove entries for specific subgenres may also include implicit information about classification, so a selection of these can also be harvested.

The musicological sources identify that there is a large number of opera subgenres and that the wording used by musicologists suggests that this high number is not always helpful. For example, Campana (2012, 202) talks about the “copious and disparate typologies” found in music dictionaries – although she only explicitly mentions one music source – and later comments on the “sheer number of generic labels” (204) which exist. Similar language is used by Carter (2014, 17), who describes the contents of the Grove typology as a “bewilderingly large number”. Doubts about every subgenre’s usefulness and necessity can also be read into these discussions (Campana 2012). The ways that subgenres are distinguished and labelled attracts attention. For example, Carter (2014, 17) describes the Grove typology as a “terminological minefield”. So, in musicological thought it can be inferred that not only are there a large number of opera subgenres, but also that this is unusual or unexpected. Moreover, the subgenres of opera are, to musicological eyes, chaotic in number and type.

It is interesting to consider how the music domain contemplates relationships between subgenres of operas. As neither the Grove typology nor the Wikipedia table contain formal manifestations of relationships between subgenres, implicit information
will be utilised instead, such as comments found in Grove entries for specific subgenres. Ideas about subgenre relationships will be identified from these music sources and then reframed in knowledge organisation terms.

Hierarchical relationships are reflected in opera subgenres. For example, the subgenre conte lyrique has a short entry in Grove, where it is described as a “term used in the late 19th century for a particular kind of opéra comique” (“Conte lyrique” 2002; italics in original). This demonstrates a genus-species hierarchy (Aitchison, Gilchrist and Bawden 2000). The subgenres of opera also present more complex hierarchies, such as polyhierarchical relationships. For example, the film musical (Traubner, Gayda and Snelson 2001) has a parent subgenre of musical, but also a parent in the genre of films. The fait historique presents a different sort of polyhierarchy. Bartlet (2002) describes it as “a type of late 18th-century French opéra or opéra comique…” (italics in original); in other words, its parent could be one of two specific opera subgenres. Hierarchically, the subgenre of fait historique as a whole has two possible parents, but each exemplar of the subgenre would have only one parent (unlike the film musical). All these examples ask questions about the quantity of levels within opera: is the fait historique a subgenre or a sub-subgenre? This questions the ontological nature of the idea of opera subgenres.

Other types of relationships are also implied. For example, in the Grove entry for commedia per musica (“Commedia per musica” 2001), the term commedia in musica is given as an alternative, which depicts an equivalence relationship. Diminutives are another example of equivalence relationships found in opera subgenres. For example, the burla is described in Grove (Burla 2001) as one type of comic Italian opera, which can have the diminutive terms of burletta and burlettina. However, the term burletta has two meanings, as it can also refer to a particular type of English opera (Temperley 2001). This example demonstrates the complexities of equivalence relationships in opera, and the importance of separating out relationships based on labelling, from relationships based on meaning.

Associative relationships are also present. For example, Märchenoper and opéra féérie are both subgenres with plots drawn from fairy tales (Millington 2001; Bartlet 2001). These two subgenres could be considered to have an associative relationship, of an undefined nature.

Figure 1 depicts the combined hierarchical, equivalence and associative relationships, using the example subgenre of Märchenoper. Some of Märchenoper’s possible relationships are shown, including its associative relationship with opéra féérie (whose two possible parents are shown via dotted lines). Note that three of the four equivalence relationships for Märchenoper, taken from Millington’s (2001) description of variant terms, appear as entries in the Wikipedia table (“List of opera genres” 2019); this highlights how some of the discussion about the quantity of subgenres (for example, Campana 2012), could actually be related to alternative titles and the instability of generic labels. So, separating out the distinguishable subgenres from mere alternative appellations can help to order the chaos. Ultimately, examining relationships highlights the complexity of opera subgenres, and shows how knowledge organisation can usefully distil and disentangle the lists of subgenres found in sources such as Wikipedia and Grove.
4.0 The classification of opera in bibliographic classifications

Considering how opera is classified in bibliographic classification schemes is critical, and it is fruitful to compare this with the music domain. This comparison is aided by utilising the idea of accords and discords, from the framework of relationships between scientific and bibliographic classifications, developed in Lee, Robinson and Bawden (2019). Eight bibliographic classification schemes are utilised for this purpose: British Catalogue of Music Classification (Coates 1960), Dickinson Classification (Dickinson 1938), Flexible Classification (Pethes 1967), Universal Decimal Classification (British Standards Institution 2006), Subject Classification (Brown 1914), Dewey Decimal Classification (Dewey et al. 2003), and McColvin and Reeves (McColvin, Reeves and Dove 1965). There is not space to reproduce a summary of the results here, but key results are identified below. Interestingly, only eight out of the 17 consulted music classification schemes are found to include any terms for specific subgenres of opera and opera-like genres (Lee 2017a).

The first important point to note is the low numbers of opera subgenres. Firstly, relatively few subgenres are listed in the eight schemes: only a total of 27 classes for opera subgenres are represented (though three classes contain multiple subgenres, to be discussed below). Looking at the schemes without opera subgenres is also fruitful: for example, Library of Congress Classification (Library of Congress 2019), is generally an extremely detailed scheme; yet, it does not choose to list categories of opera nor separate opera from other musical-dramatic works (Library of Congress 2019). Interestingly, some schemes state their mistrust of opera categorisation explicitly: for example, the Expansive Classification (Cutter 1891-1904) and Olding’s (1954) classification both state that they do not consider dividing opera into subcategories to be useful.

This is in direct contrast to the music domain where the large numbers of subgenres was a focus point, and hence shows discord between bibliographic classification and the music domain. There are a number of possible explanations. First, the main rationale for the bibliographical classification schemes is retrieval; so, while many opera subgenres may exist, there may not be warrant for their inclusion in a bibliographic classification scheme. Second, the complexities and bewilderment commented upon by the musicologists might lead to lack of standardisation in subgenres, which in turn leads to a lower probability that subgenre information is useful to users. Third, the discord could be a reflection of the shallower levels of information seen in bibliographic classification schemes than domain-based classifications. However, the lack of opera subgenres in bibliographic classification schemes which are notoriously detailed (for example, Flexible Classification and Library of Congress Classification) suggests this is not the only (or even primary) explanation.

The eight bibliographic schemes also reveal a distinct lack of coherence between themselves: out of 27 classes, 14 appear in only one of the eight schemes. While six of these “single-appearances” could be explained away as they are from a notoriously detailed scheme (Flexible Classification), the other five cannot. The lack of coherence

---

1 The Library of Congress Genre/Form Terms provides a potential additional source for the bibliographic classification of opera subgenres. However, a cursory glance through the variant terms attached to the entry for opera (Library of Congress 2020) shows equivalent or less detail than the bibliographic schemes, so has not been contemplated further within the space limitations of this paper.
could be considered a realisation of classification chaos, perhaps enhancing the views of the Olding and Expansive classifications about the foolishness of trying to categorise opera.

Fourteen subgenres appear in the combined bibliographic schemes which do not appear in the Grove typology or Wikipedia table, showing further discord between the music domain and bibliographic classification. In some cases the bibliographic classification scheme examples are more detailed than the music domain: for example, great operetta and small operetta appear in a bibliographic classification scheme but not in the Grove typology or Wikipedia table. This weakens any argument that the lack of subgenres in the bibliographic schemes is due to lack of detail. In other cases, the bibliographic classification schemes have broader categories which do not refer to specific subgenres but could be considered as broad types of opera – for example, light opera and comic opera. This suggests a domain/bibliographic discord in the idea of opera’s units.

The bibliographic classification schemes demonstrate some explicit relationships between subgenres. For example, Flexible Classification (Pethes 1967) has a hierarchical relationship in operetta, where the sub-classes of great operetta and small operetta have the class operetta as their parent. A form of equivalence can be seen in the use of combined classes for subgenres, such as McColvin and Reeves’s (McColvin, Reeves and Dove 1965) shared class for light opera, musical comedies and revues. So, there is some accord between bibliographic schemes and the domain, in that hierarchy and equivalence relationships are present, albeit with different levels of implicit-ness.

5.0 Towards a model of classifying opera subgenres

There is, however, another way of contemplating the classification of opera subgenres, away from their inter-subgenre relationships: consider the categorisation of the attributes of each subgenre. Sources in the music domain comment on this as a categorisation method. For example, Campana (2012) and Carter (2014) remark upon the different ways that subgenres are delineated in the Grove typology, though neither author is intending to provide a complete list of the distinguishing features found in the typology or from a theoretical perspective. Campana’s (2012) and Carter’s (2014) combined list of attributes include formal qualities (including the interrelation between speech, music and dance), subject matter, medium (in this case meaning the media of performance, such as radio or television), function (for example school operas), national operas and subgenres (relating to the idea of place) and a sense of historical period. (Note that as medium has another meaning in music classification (Lee and Robinson, 2018), the term “dissemination” will be adopted instead for the category containing foci such as television or film.) At this juncture it is useful to revisit Dahlhaus’ (1987) general list of a genre’s constituents; this sees some overlap (form and function), and also adds the ideas of medium (who is playing and performing) and text. If we were to argue that what distinguishes one subgenre from another is the same idea as what constitutes that subgenre (or genre), we can combine both sets of factors. Therefore, we could see that loosely speaking, opera has (at least) eight constituents, which translated into standardised music classification terms, are as follows: form, subject, dissemination, function, place, time, medium, and text.
Note that ideas such as place are complex in opera categorisation. Place can represent the boundaries of a subgenre’s world, its germination or its association. Furthermore, attributes do not always work independently. For example, there is a nebulous boundary between nationality and place, as hinted at in Carter’s (2014) depiction of “national genres”, and place can be associated with text via language. Ultimately, this categorisation of attributes is useful, but does not always distinguish between single subgenres: for example, opera buffa and burla both describe Italian comic operas of the 18th century. Put simply, categories are invaluable for studying shared properties of opera subgenres; however, they cannot always elide between one subgenre and another and cannot explicitly track genre development.

So, a model is proposed in Figure 2, which visualises opera subgenres as both a system of relationships between individual subgenres and as the categories of information which constitute individual subgenres. The model has two planes: the inter-subgenre plane and the categorisation plane. The inter-subgenre plane allows for the complexities and quantity of subgenres by disentangling the web of relationships between subgenres; whereas the categorisation plane shows how each subgenre contains categories of information, which could be seen as working in tandem with the subgenre-to-(sub)genre relationships found in the inter-subgenre plane. Links between the categorisations and music’s facets could be perceived, between categorisations which also appear as meta-facets (Elliker 1994) such as function, place, time and medium. (Form is, of course, already part of the form/genre facet (Elliker, 1994; Lee, 2017a)). These constituent-facet connections are demonstrated with freely-drawn blue arrows, representing the informal nature of these relationships. Finally, the associative relationship in the inter-subgenre plane could be concurrently viewed as a connection between subgenres which share particular constituents, especially for operas sharing the same subject material. This is also depicted with a freely-drawn blue arrow, representing the intangible nature of this connection. Therefore, this model demonstrates how knowledge organisation can be used to unpick and provide order to the music domain’s multitudinous collection of opera subgenres.

6.0 Concluding thoughts

Opera subgenres are viewed as being somewhat tumultuous from the perspective of the music domain. This paper has analysed the opera subgenre soup, in order to unpick what is happening. Using a knowledge organisation framework, it can be seen that at least some of the superfluous number of subgenres might be explained by the presence of the same subgenres appearing multiple times in resources with different labels. Furthermore, the subgenres of opera can be better understood as a complex web of different relationships, rather than through a one-dimensional list. Lastly, the two-plane model suggests that there are connections between categorising the inter-subgenre relationships and categorising the information which informs the delineation of those subgenres. This is a novel way of considering the classification of (Western art music) genres, and could be utilised to examine other knotty sets of genres.

Exploring the bibliographic classification of subgenres introduced some interesting ideas. Tangible discords with the domain are shown: the quantity of subgenres differs, as does the level of hierarchy represented by classes in some cases (both broader and narrower). These cannot be explained solely by the retrieval-focused nature of
bibliographic classification schemes. However, the findings can also be read as accordance between scientific and bibliographic classifications: the lack of coherence in the bibliographic scheme’s categorisation of opera subgenres could be seen as a reflection of the confused and unruly set of opera subgenres found in the music domain. This paper is a preliminary step in furthering understanding of Western art music genre classification, and so future research could see a similar analysis applied to other Western art music genres. Furthermore, it would be productive to see how the results from this paper fit into genre classification research pertaining to other music traditions. It would also be interesting to contemplate other sorts of artistic works using the two-plane model. Consequently, this paper illustrates how knowledge organisation can provide order to operatic chaos, and in the process, advance our understanding of music classification and knowledge organisation more generally.

Figure 1. The subgenre relationships for Märchenoper

Figure 2. Model of the classification of opera subgenres
References


The Bias in Ontologies
An Analysis of the FOAF Ontology

Abstract:
Knowledge Organization Systems (KOS), like thesauri, classification schemes, taxonomies, or ontologies, are essential tools for the organization and representation of information in various contexts and are often understood as neutral tools without any bias. However, we can argue representing information, even unconsciously, we can describe some form of prejudice, that is, what is a bias, by the person who creates the system. This selection of elements represented is required in any KOS since every representation has a specific function that is related to a context. Ontologies are an excellent example of this because, as Guarino, Oberle, and Staab (2009) state, these KOS need to delimit their goal to enable reuse and avoid problems arising from excess of ontological commitment. With that in mind, we seek to discuss possible bias that a KOS may have, but focusing on ontologies and taking as our object of analysis the Friend of a Friend (FOAF) ontology. Thus, we characterized this research as descriptive, with a qualitative approach. The objective of the review is to understand the implications of bias in these KOS, also seeking to discuss how Knowledge Organization, as a field of study, can act in the development of tools that recognize its own bias and still be able to perform its functions. For the analysis, the theoretical framework of Discursive Semiotics is used, which studies the formation of meaning as a phenomenon from a model called Generative Trajectory of Meaning (GTM). From this perspective, we can understand bias as a product of semiotic processes – figurativization, thematization, and discursivization (Greimas and Courtés 2013) – involving the KOS developer social-cultural contexts (Gomes and Barros 2019a, 2019b). From this theoretical understanding, all the elements that constitute the FOAF ontology – classes and properties – are analyzed, as well as its documentation available online. We concluded that bias is an inherent feature of a KOS and that Knowledge Organization could focus on conducting studies on technologies that enable information retrieval, taking into account this aspect of its tools. In order to: (1) go beyond the KOS bias, using, for example, "see also" connections that act as hyperlinks to systems with other biases that best fit the user's needs; or (2) "learn" the various perspectives that exist on the same topic, represent them in a KOS and drive users to those best suited to their needs – in which case issues such as Machine Learning and Artificial Intelligence should enter the discussion, making this tools more semantic enriched.

1.0 Introduction
This study aimed to discuss the consequences of the presence of one or more biases in Knowledge Organization Systems (KOS), taking ontologies as an example of the presence of this bias. For this, we analyzed the elements that make up the Friend of a Friend (FOAF) ontology, which aims to represent individuals and their relationships, as its name already implies.

We also seek, from the analysis, to demonstrate the importance that understanding and explaining the bias of a KOS can have since considering it can be fundamental for efficient information retrieval to occur.
2.0 The method

For the analysis of the FOAF ontology, we used the theoretical tools of Discursive Semiotics, considering that, based on this theory, it is possible to understand the formation of the meaning of a discourse, which, in the present case, is an ontology. By discourse, we mean the concretization, in language, of a particular social, historical, ideological, and environmental context (Possenti 2009). Thus, Discursive Semiotics studies the mechanism by which a given discourse is shaped, and when applied to ontologies or KOS in general, it can reveal some important aspects of its constitution, especially those related to aspects that shape it (Gomes and Barros 2019b).

Discursive Semiotics, for didactic reasons (Greimas and Courtés 2013), adopts a model called Generative Trajectory of Meaning (GTM), which organizes the formation of discourse in two levels of depth – that is, it goes from the semantic level to the discursive. It is essential to highlight that there are several GTM models, and we adopted in this work the one developed by Greimas and Courtés (2013).

The GTM has two aspects: (1) semionarrative structures and (2) discursive structures. A semiotic analysis starts from the discursive structures – which organize the contextual elements that form the discourse and make it understandable – and go into the semionarrative structures – formed by elements called actants, which can act to each other and transforming it. The actants, in turn, are formed by even smaller and completely abstract units, called semes, which gain meaning from their interaction with opposite, complementary and contradictory semes.

The interaction between semes generates what we can understand as the "meaning" or “particular meaning” of a given the word.

As the focus of this work is the bias that possibly exists in an ontology, we chose to pay more considerable attention to the GTM's discursive level. At this level, there is a series of operations, which take place from semionarrative structures, covering them with the contextual component mentioned above – the social, historical, ideological, and circumstantial context. Operations that occur at this level are:

- Discursivization: it makes explicit those involved in the discourse, forming actors (actorialization), as well as space (spatialization) and time (temporalization) in which they were enunciated;
- Figurativization: the actors gain a semantic investment, becoming figures; that is, they we can understand as something real.
- Thematization: it is an abstract thematic covering on which the figures act.

Thus, all operations that occur from semionarrative to discursive structures come into contact with some linguistic system, thus forming lexemes (in a way, words), from which we interact and put the discourse into practice, considering a given context. However, this explanation is still too general and was designed especially for discourses in action, which is not the case with ontologies. These types of KOS were built generally, because of their reuse (Gruber 1995) and the domain representation and, because of that, ends up being quite a generalist.

In previous works (Gomes and Barros 2019a), we highlight that the concepts that constitute ontologies can be understood, like any word, as lexemes formed from the semiosis operations explained above. Thus, the concepts can be considered figures within the represented discursive universe; that is, they went through the process of figurativization so that they have a precise semantic coating. Thematization occurs based
on the domain itself that is being represented in the ontology, considering that the understanding of the concepts is only possible from the abstract coating given by the themes. Finally, based on the existence of the ontologist (after all, it is they who constructs the discourse, the ontology), it is possible to affirm that actorialization, temporalization, and spatialization also occur. These processes allow us to situate concepts from the referents they seek to represent, together with the perspective of those who produce the discourse. The following figure shows how the elements that make up an ontology (including its developer) also in light of the level of GTM's discursive structures:

![Diagram](image)

**Figure 1 – The GTM’s discursive structures on ontologies**

Therefore, the discussion carried out in this work we based in the idea that lexemes form an ontology gain meaning from the performance of a series of semiotic operations involving the formation of figures – which gain meaning because they are linked to some theme – and the formation of the ontologist as a semiotic actor, present in a specific time and space.

Thus, the study of the FOAF ontology started from the semiotic approach. This research we characterized as descriptive and qualitative. The FOAF ontology was analyzed from its documentation¹, which explains its objectives and constitution (classes and properties), and used the concepts presented in this section as a theoretical foundation.

The analysis sought to clarify how the semiotic processes that occur in the GTM discursive structures end up generating a bias in the ontology, even against the will of the ontologist. Initially, we studied the general information of FOAF, such as its objectives and used, then we move on to the study of the classes and properties that comprise it.

To check for the presence of bias in the ontology, we observe which of the classes and properties, as well their descriptions in the documentation, dialog with a more specific context or ideology – for example, which classes or properties represent things

that are present in a given region of the world. Therefore, a class like *foaf: Person*\(^2\), based on the distinction between what is and is not a person, says much less about the bias of ontology than a property like *foaf:gender*, which has more significant social implications, given the discussions about gender, so we chose to pay more attention to *foaf: gender*. Thus, we selected some of these classes and properties to guide the discussion about the ontology bias.

Starting from the explanation of the bias present in FOAF, we propose a discussion about how this bias can affect an ontology and the actions that can take so that this phenomenon is not necessarily an intentional problem.

### 3.0 The FOAF ontology and its bias

The FOAF ontology, as stated in its documentation, was developed to connect people and information through the Web, and this information can be anything from documents of any support, data, or even just ideas in someone’s mind. For this, FOAF integrates three types of networks:

“social networks of human collaboration, friendship, and association; representational networks that describe a simplified view of a cartoon universe in factual terms, and information networks that use Web-based linking to share independently published descriptions of this inter-connected world” (Brickley and Miller 2014)

FOAF’s terms were divided into three broad categories: (1) Core, formed by terms that involve people and groups regardless of time and technology; (2) Social Web, formed by terms related to activities carried out on the Web; and (3) Linked Data utilities, formed by terms that can be useful for the Web community to connect data. Despite this division, the documentation explicitly distinguishes only the terms of categories 1 and 2. It is worth noting that this ontology can always be updated, with the insertion of new terms and that old terms, called "archaic" in the documentation, are always maintained in order to enable old forms to become modern again (Brickley and Miller 2014).

The following image, taken directly from FOAF’s documentation, explains all its classes and properties:

**Figure 2 – FOAF classes and properties**

As we can see, some classes and properties are quite general, as is the case with *foaf:Agent*, defined as something capable of doing something. This class can be used to represent situations in which the being who is acting in a given situation is not exactly a person or group of people (it can be a software bot, for example). A subclass of

---

\(^2\) To reference the names of the classes and properties of the ontology, we have chosen to use the same form as the one present in the FOAF documentation. Thus, something like *foaf: Agent* (capitalized) is a class, whereas *foaf: knows* (lowercase) is a property.
foaf:Agent is foaf:Person, used to represent people, and they may be alive, dead, or not even exist at all. This broader scope is fundamental in the case of FOAF, which is an ontology that aims to be widely reused in the most diverse situations that involve the connection between people and information on the Web.

In the theoretical scope of Information Science, a concept present at FOAF that generates much discussion is represented in the class foaf:Document. We know that there are several perspectives on the concept of a document, such as those presented by Suzanne Briet (1951), Michael Buckland (1997), and Berndt Frohmann (2009). The FOAF documentation states only the following about this class: “The Document class represents those things which are, broadly conceived, ‘documents’” (Brickley and Miller 2014). With this definition of the class, quite broad, it is already possible to perceive more clearly the presence of bias in the ontology, although that specific bias does not generate any negative consequences. The definition of a document is not a common concern in all areas, but it is crucial for Information Science, which has, in this object (Buckland's "information as a thing" (1991)), one of its focuses of study.

Thus, the ontology, even if indirectly and without an intention, ends up “taking sides”, even if this does not produce harmful effects for what it proposes. However, there are elements in FOAF that are less subtle, and that makes the ontology bias even more explicit.

The foaf:gender property, already mentioned, can cover several different perspectives. In a more conservative ideological perspective, there are only two genders, male and female; however, in a more progressive perspective, gender is understood in a much less fixed way. The FOAF documentation says the following about that property:

“The gender property relates an Agent (typically a Person) to a string representing its gender. In most cases, the value will be the string 'female' or 'male' (in lowercase without surrounding quotes or spaces).

Like all FOAF properties, there is, in general, no requirement to use gender in any particular document or description. Values other than 'male' and 'female' may be used, but are not enumerated here. The gender mechanism is not intended to capture the full variety of biological, social, and sexual concepts associated with the word 'gender'” (Brickley and Miller 2014).

In other words, foaf:gender recognizes the diversity of perspectives about gender, so to avoid taking a stand and maintaining a high level of generality in ontology, they choose to leave the term open, without mandatory use in any circumstances, but indicating that the most common way to fill this property is with the strings "male" and "female." The authors themselves make it clear, later in the document, that they are aware of the difficulty of working with the concept of gender: "We have tried to be respectful of diversity without attempting to catalog or enumerate that diversity" (Brickley and Miller 2014).

As explained in the previous section, we can explain this phenomenon of bias in an ontology using Discursive Semiotics and taking into account that an ontology is a discourse, the person responsible for developing the ontology, in the case of FOAF, ontologists have a bias. These subjects, being inserted in a given spatial, temporal reality, and being able to act on the things that exist in the world, end up transferring to the ontology a section of all these aspects that shape them. To understand this, we could imagine how the description of foaf:gender could be different 30 years ago, that is, the current description is the result of the temporalization process explained previously, responsible for inserting the actors in a particular temporal reality.

The spatial aspect of FOAF’s discourse can be seen when comparing the pairs of properties foaf:firstName X foaf:givenName and foaf:lastName X foaf:familyName. In
a way, each pair refers to the same thing, however, as the developers themselves claim, the “concepts of ‘first’ and ‘last’ names do not work well across cultural and linguistic boundaries; however they are widely used in address books and databases” (Brickley and Miller 2014), that is, although the two ways of referring to someone’s first and last names are valid in some cultures, in others someone’s last name is not the same as their family name. For example, in some Eastern countries, the first name is the family name, and the last name is the given name.

We can also point out, as another example of the presence of contextual aspects of ontologists at FOAF, the inclusion of the property, labeled as “archaic”, foaf:dnChecksum (which could be used to verify the data integrity in the DNA transfer from a person), created as a joke by the developers. The only objective with this inclusion was to demonstrate the great diversity of properties that could be created to identify someone, some of which, the developers add, “we might find disturbing” (Brickley and Miller 2014).

4.0 The consequences of the presence of bias in FOAF and KOS in general

In order to understand how a specific bias can affect an ontology like FOAF, it is worth highlighting some requirements that it must meet. According to Gruber (1995), an ontology must have:

- clarity: the concepts present in an ontology must be clear and objective so that the definitions do not depend on social contexts or computational requirements. That is why ontologies are generally developed from a formal language, using logical axioms. Also, in order to facilitate the understanding of ontology by a human being, it is highly recommended that the definitions be documented in natural language (as is the case with the documentation analyzed here);
- coherence: the axioms that make up the ontology must be coherent so that those logical inferences can be easily made. There can be no contradictions between the definitions;
- extendibility: an ontology must be developed, taking into account that the vocabulary can be reused in some other situation. Thus, the elements that make up the ontology must be open enough so that new terms are inserted without the need to change those already present;
- minimal encoding bias: ontologies must be formed by the concepts they want to represent regardless of the computational language used in their development since they are used by different systems of representation and styles of representation;
- minimal ontological commitment: an ontology must have a minimal ontological commitment as possible in order to be able to share knowledge and reuse it, in addition to interoperability between systems. The fewer statements about the discursive universe that are made, the better, so that it is preferable, in many cases, that only the necessary (but not sufficient) characteristics of a given concept are made explicit (however, for reasons of clarity, when possible, a complete definition, with necessary and sufficient characteristics, must be provided).

An ontology that “take sides” in a very explicit way faces the risk of not being able to meet its objectives efficiently, as this may affect some of the above requirements,
more specifically its clarity, extendibility, and ontological commitment. For example, if FOAF adopted only `foaf:firstName` and `foaf:lastName` to express the names of individuals, it would be committing itself to a more specific reality, unlike what happens with `foaf:givenName` and `foaf:familyName`, which serves the largest share of the world population. Besides, in the case of `foaf:gender`, the decision not to make this property mandatory or to define default values for it also had the purpose of maintaining the above requirements, which would not happen if the developers were more assertive in their opinion.

In the case of other KOS, such as thesauri, the requirements above are different, especially concerning clarity or even ontological commitment. Many KOS must represent a given domain of knowledge in the most transparent possible way, with precise definitions (very different from the general characterization of ontologies). In such cases, the presence of a bias can further affect the retrieval of information through this KOS, as it can affect the way users interact with the information they seek (that is, KOS influences how the user understands the information) or negatively affect the retrieval of information, as how the information was represented may not be consistent with what the user understands. An excellent example of this can be seen in the work of Miranda and Costa (2019), in which the authors analyze the way a bibliographic representation of texts from Umbanda is made, finding cases in which books referring to this religion were inserted in class 133.4 of the DDC system (Demonology and witchcraft). In this situation, the bias of those responsible for classification is quite evident, since, for Umbandists, their belief has nothing to do with witchcraft.

Based on these examples and the semiotic process of meaning formation not only of ontologies but of KOS in general, we understand that bias is an inherent element in information representation, although we try to avoid it. Even a very generalist ontology such as FOAF ends up “taking sides” regarding some themes.

5.0 Conclusion

In this work, we started from the principle that a KOS can be understood as a discourse, being able to be analyzed by the theoretical tools provided by Discursive Semiotics, which allows us to understand how the meaning of that discourse is constructed. Thus, we analyzed FOAF ontology as a discourse, which carries with it some contextual aspects regarding who develops it (who they are, where they live when they are developing the ontology), so that, even against their will, it ends up transmitting in their elements some of these aspects. In other words, FOAF, like any KOS, is biased, and this is natural.

However, it is necessary to think about how Knowledge Organization can deal with this bias since a KOS that transmits a particular idea in a very understandable way can end up negatively affecting information retrieval from its use.

The inherent bias in KOS, we believe that to deal with this situation, it is necessary to develop studies focused on technologies that enable the use of bias in favor of the users. An idea can be found in FOAF itself:

“If people publish information in the FOAF document format, machines will be able to make use of that information. If those files contain ‘see also’ references to other such documents in the Web, we will have a machine-friendly version of today's hypertext Web” (Brickley and Miller 2014).
The use of “see also” references in KOS, in general, could be a way of connecting systems that deal with the same topic but have different perspectives. Alternatively, even within a single system, to support the different perspectives that may exist.

Another possible solution, considerably more complex, would be the KOS itself, being part of a more extensive system, having access to what the user tends to search for, and directing search results to those already known needs. In other words, KOS would learn what biases might exist in any topic and make a comparison between what it knows and what its user knows or usually research. In this case, the Knowledge Organization should bring issues such as Machine Learning and Artificial Intelligence to the discussions.

Thus, the discussion proposed in this paper aimed to explain how, even in generalist KOS like ontologies, a bias can be evident, which can affect information retrieval. As this bias is natural and attempts to avoid it are not entirely adequate (for reasons explained by semiotics), we highlight that a paradigm shift towards accepting the KOS bias as a way to benefit the user could bring benefits. Since we understand bias as inherent to KOS, this work proposes that the developers of these tools pay attention to their ideas and recognize the wide variety of existing perspectives on the theme that they seek to represent in their systems. With this in mind, they could look for ways to direct users to tools that best fit their needs. This paradigm shift could occur from more straightforward changes, such as the “see also” references, or in the depth of studies on technologies that allow the automation of this process of verifying the bias and directing the user.

References


A System for Specifying Semantic Relations for Knowledge Representation

Abstract:
Semantic relations are fundamental for understanding the nature of the connection between two concepts in a domain. This paper presents a model for extracting semantic relations for the representation of knowledge from academic documents in the context of the Portuguese language. A Web information system called Semantizar was developed to support the extraction of semantic relations from classificatory structures that represent specific academic documents. To evaluate the qualitative performance of Semantizar, a case study was carried out, which pointed to important contributions to research about semantic relations extraction. According to outcomes, when two concepts of a classificatory structure exist in a sentence, a semantic relationship between them can actually exist. Finally, it is concluded that this research is relevant because it brings important findings for the extraction of semantic relations for the knowledge representation of academic documents to be applied in the Brazilian scenario.

1.0 Introduction
Semantic relations are fundamental for understanding the nature of the connection between two concepts in a domain. According to Khoo and Na (2006) and Green, Bean e Myaeng (2011), concepts can be seen as blocks of knowledge, and relationships are links that connect and hold these blocks together within the structures of knowledge in people's minds.

Some factors can influence the specification of semantic relations, including language and culture. According to Khoo and Na (2006), it is difficult to analyze the meaning of concepts and their relations when they are taken apart from language because each language has its characteristics and is linked to the cultural factor. According to Khoo and Na (2006), it is difficult to analyze the meaning of concepts and their relations when they are taken apart from language because each language has its characteristics and is linked to the cultural factor. Therefore, in the literature review conducted by Maia (2018), was identified the lack of research concerning the theme in Brazil.

Subsequently, a Semantic Relations Extraction Model was formulated. This model’s goal is to be a theoretical construction of a way of establishing relationships between concepts from academic document, in order to create semantic structures. The model was expanded to a computational prototype called Semantizar.

2.0 The Semantizar
The following procedures were observed in the development of Semantizar: (1) specification, (2) data modeling, (3) architectural design, and (4) prototype implementation in a Web system. Of these, the specification and the prototype implementation will be presented in this paper.

In the first stage of the development of Semantizar - specification -, was made the description of the algorithm for the Semantic Relations Extraction Model, which considers as inputs a classificatory structure and its respective academic document, from which the structure originated. The classificatory structure is broken down into concepts and the academic document is decomposed into sentences.
After this decomposition into concepts and phrases, the Semantizar scans all the phrases searching for pairs of concepts. In the first scan, the pair of concepts 1 and 2 is enhanced, then is checked their existence in each sentence until the last one, regardless of the discovery of a pair of concepts in a sentence. In this way, the search is done throughout the document. If the concept pair is found in a sentence, that sentence is highlighted, so that a manual check confirms or denies the existence of a semantic relationship between these concepts. Then, the Semantizar combines the concept 1 with all the other concepts of the classificatory structure, checking whether each combination exists in a sentence. Upon ending the combinations with concept 1, the Semantizar subsequently makes combinations with concept 2 and checks whether they exist in all sentences. In this way, the Semantizar makes combinations of pairs of concepts with all the concepts of the classificatory structure and checks the existence of each of these pairs in each sentence of the academic document, from the first to the last one, as illustrated in Figure 1.

Figure 1: Search iterations of pairs of concepts in the sentences of the academic document.

The prototype implementation phase was divided into three activities: (1) data input, (2) reading and preparation and (3) extraction of semantic relations. The Figure 2 shows the initial interface of Semantizar prototype.

In the data input activity (1), the user informs the metadata of the academic document from which intends to extract semantic relations, and then sends the files of the publication and classificatory structure. The publication being a .pdf file and the classificatory structure being a plain text file with .txt extension.
In the subsequent activity, reading and preparation (2), the Semantizar checks whether each term of the classificatory structure exists in the database. If the term does not exist, it is automatically registered by the system in this way: The PHP programming language, chosen for the implementation of the model, allows text files to be converted into vectors. In this sense, the file that refers to the classificatory structure is automatically converted into a vector of terms, in which each line of the structure (which refers to a term) is transformed into a position of the vector. Therefore, the algorithm goes through each position of the vector, checking if the content of the position, which is the term of the classificatory structure, exists in the database; if it does not exist, the term is automatically registered by the system as a noun. This was established because, grammatically analyzing, the terms of the classificatory structure denote the nouns.

The second task of the reading and preparation activity is the preparation of the publication file (which is an academic document: a thesis or a dissertation) for the manipulation that will occur in the next activity of the prototype. Due to the programming language chosen for the implementation, it was necessary to convert the text in .pdf format to a temporary file in .txt format.

The activity of extracting semantic relations (3) is considered the most important, being the core of the Semantic Relations Extraction Model. It consists of two tasks: the first searches for pairs of terms from the classificatory structure in sentences of the publication to which the structure refers. The temporary file of the academic document created in the previous step is transformed into a string (variable that stores alphanumeric characters). Subsequently, this string is decomposed into a smaller string each time a period (.) is found in the file. In this way, each position of the vector is a sentence of the publication, separated by a period. Then, the sentence vector is scanned seeking terms of term vector in each sentence vector position. If a term is found in the sentence, the Semantizar goes through the other positions of the vector of terms checking if there is another term of the structure in the same sentence. In affirmative case, the sentence is taken to compose the interface created for the user to validate the semantic relationship.

The validation of the semantic relation is the second task of the relation extraction activity. If the user agrees that there is a semantic relation between the two concepts of...
the classificatory structure found in a sentence in the publication by the Semantizar, (s)he is directed to semantic relation register interface.

In the register interface the user specifies the semantic relation according to judgment when analyzing the sentence. Besides that, the user determines the type of semantic relation, the inverse relation, if any, and points out the properties: symmetry and reflexivity. The transitivity property was not considered because it is understood that it applies to ternary relationships, which is not the case in this paper.

The types of semantic relations that appear in Semantizar were the result of a literature review by Maia, Lima and Maculan (2017), which elaborated a taxonomy with 63 types of semantic relations classified as hierarchical, equivalent and associative.

3.0 Case study

To evaluate the efficiency of Semantizar in the extraction of semantic relations, a case study was carried out. It was organized according to the Experimentation Process methodology proposed by Wohlin et al. (2014), which considers three stages: (1) definition of the case study, (2) planning and (3) operation.

In the first stage, definition of the case study (1), it was determined: (a) the object of the case study, which is the semantic relations; (b) the objective: to verify the efficiency of Semantizar; and (c) the context, which are theses and dissertations in the domain of Knowledge Organization and Representation.

After the case study definition phase, the second stage, planning (2), follows. According to Wohlin et al. (2014), this stage basically indicates “how” the case study will be conducted. Thus, in planning, it was established: (a) the sample and (b) the analysis of the data to be collected.

For the sample, was chosen the faceted structure of MHTX, by Lima (2004) and the thesis Fatores Interferentes no Processo de Análise de Assunto: Estudo de Caso de Indexadores (Interfering factors in the process of subject analysis: indexer case study), by Naves (2000). The MHTX is an in-context hypertextual navigation model to organize theses and dissertations, aiming to support the reading and retrieval of these documents in Digital Libraries of Theses and Dissertations. In the MHTX prototype three navigation tools was created: the expanded summary, the concept map and the faceted structure. In its implementation, the Naves’s (2000) was used aforementioned thesis to instantiate the tools created. Among these tools, the faceted structure presents the characteristics of the desired sample for Semantizar.

As mentioned in the planning stage, in addition to defining the sample, the analysis of the data was determined. In this case, was decided to perform quantitative and qualitative analyses in order to identify: (I) the number of semantic relationships suggested by the application due to the amount of semantic relations that actually exist (this factor is important in order to evaluate whether the Semantizar has the potential to automatically extract semantic relations); (II) the concepts that are most likely to be semantically related (this parameter can point to the key concepts of the analyzed publication); (III) the characteristics of the semantic relations found; and (IV) the parallel between concept relationships in the original faceted structure and the resulting representation from Semantizar.

Following the case study process, the next step is the operation (3). This phase consists of the execution of the previously defined and planned case study (Wohlin et al.
2014). For this, three procedures were necessary. The first was preparation, which involved the clipping of the sample from the classificatory structure and the publication, selecting subjects from the Personality facet, as can be seen in Figure 3. In this case, the terms idea, thought and concept were broken down. Regarding the academic document, we decided to consider chapters 2, 3 and 4 of Naves’s thesis (2000). This choice was due to the fact that these chapters constitute the conceptual definitions in the thesis in question. Figure 4 shows the thesis’ summary, in which these chapters can be seen. We also decided to remove images from the chapters, since Semantizar cannot support image analysis. The second procedure, execution, comprised the processing of the sample snippets on Semantizar. Finally, the last procedure was the validation, in which a refinement and compilation of the data collected was made in order to avoid interferences in the data analysis and interpretation. Both in the validation of the data and in the results, the concepts were observed individually and with their pairs.

![Figure 3: Snippet of MHTX faceted structure](image1)

![Figure 4: Fragment of the expanded summary of the Naves (2000)](image2)
4.0 Results

Figure 5 presents a conceptual map generated from the classificatory structure without Semantizar. In this Figure, it appears that the explicit semantic relations are due to the naming of the sub-facets used by Lima (2004) (see these sub-facets highlighted in Figure 3). Also, it is observed that there is a semantic relation between information professional (Profissional da Informação) and librarian (Bibliotecário) created from indentation that denotes a type of hierarchy. However, it was not possible to specify what the semantic relation really is.

![Figure 5: Clusters resulting from conceptual map of the MHTX faceted structure without Semantizar.](image)

In the organization of the faceted structure in the concept map, the presence of two clusters is verified, as highlighted in Figure 5. The first group consists of concepts related to text (texto), and the other consists of concepts related to indexer (indexador). In clusters, objects belonging to a group are related to each other, however, they do not relate to concepts that are outside of their group. Therefore, in the conceptual map generated from the semantic relations found in Semantizar, there is a cohesion between all concepts in such way that these clusters are not possible to be obtained, that is, the concepts are all related to each other, as seen in Figure 6. With the use of Semantizar, it was possible to explicit 101 semantic relations.

5.0 Contributions

The relations between all concepts were possible with the support of Semantizar, which allowed the creation of a representation that covered every concepts of a semantically related classificatory structure. So, the user can view all possible relations between the concepts. Therefore, Semantizar achieved its goal of semantically enriching a classificatory structure.

The performance of the case study, within the scope presented, was possible due to the computational support of Semantizar. This task performed manually could demand time and effort on the part of the professional who executes it. In this sense, Semantizar facilitated the extraction and explanation of semantic relations, essential for the knowledge representation, semi-automating this task. Thus, the Semantizar contributed to the knowledge representation based on a classificatory structure, showing itself to be objective when detecting two concepts in a sentence from extensive text. The identification of concept pairs within sentences is one of the most laborious steps in the context that Semantizar was created to operate.
Therefore, through Semantizar, it was possible to establish 101 semantic relations, including inverse ones, in 53 different concept pairs, based on a classificatory structure containing 22 concepts. With that, it was possible to improve the semantics of the sample.

Another important contribution of Semantizar was that, in a way, it acted as a validating agent of the classificatory structure, as it indicated 199 signs of semantic relations in 131 sentences from the snippet of the academic document from which the classificatory structure originated. It is believed that this amount of evidence is an important indicator to suggest that the structure relevantly represents the academic document. Another important factor in this regard concerns the sentences in the academic document, as some of them encompassed more than a clue of semantic relation, which reinforced confidence in the classificatory structure.

In addition, it was possible, through Semantizar, to indicate the most important concept pairs for the academic document. During the analysis of the data, it was found the concept pairs that occurred most frequently were the most relevant in the represented domain. In the case of the sample, two pairs stood out: indexer and document (indexador e documento) and indexer and text (indexador e texto). Considering these pairs were extracted from the thesis entitled Interfering factors in the process of subject analysis: indexer case study, by Naves (2000), it can naturally be said, even without reading the entire content of the academic document, that such pairs are the most important in the
domain. Consequently, Semantizar also indicated the most important concepts for the academic document. In the same way as for the pairs of concepts, considering the concepts that occurred most frequently in the semantic relations, it can be said that the most important ones in Naves’s thesis (2000) were indexer (indexador), text (texto) and document (documento), also taking into consideration that text and document were classified as almost synonyms, which also indicates coherence in determining this semantic relationship.

It was also found during the analysis that, many times, the inverse relations were not possible because the relations were indirect, the concepts existed: subject (assunto), content (conteúdo) and information (informação). Therefore, it was discovered that these concepts should compose the classificatory structure due to the fact that they repeatedly occurred and that they are representative for the domain. Similarly, it was noticed the concepts with the most evidences false in the context in which they were found were those that are routinely used in academic documents, such as concept (conceito), idea (ideia) and authors (autores). In this sense, a review of the classificatory structure to indicate whether they remain to represent the academic document was suggested. Thus, in pointing out these suggestions, it can be said that Semantizar can operate to refine the classificatory structure.

It was also noted that the extraction of concepts can be performed from lists of terms, as the inherent hierarchy of the classificatory structure was not decisive in Semantizar for the indication of the existence of a semantic relation.

Finally, the case study carried out made important contributions to research on the extraction of semantic relations. Those contributions are: 1) The presence of two concepts in a sentence is an indication of the existence of a semantic relation between these concepts; 2) A pair of concepts can have more than one semantic relation; 3) A pair of concepts can have the same semantic relation, even in different contexts; 4) The context, and the knowledge about it, is fundamental for determining the type and/or subtype of the semantic relation for the same pair of concepts with different relations; 5) The determination of semantic relations, in the way it was treated by Semantizar, depends on human interpretation; 6) The verbs are the main grammatical class for defining a semantic relation and; 7) Not all semantic relations can be explained.

References

Representing Faceted Classification in SKOS

Abstract:
Faceted classification is one of the most important contributions from knowledge organization to information resources management. Proposed in the twentieth century by pioneers such as Bliss and Ranganathan as an alternative to the rigidity of enumerative classification, faceted classification has gained importance, both academic and practical. Today it is widely used in web architecture, from scientific to commercial sites, and even as a methodology for the development of ontologies. The web is evolving towards the vision of the Semantic Web, in which resources are not only navigable and understandable by humans, but also with content that has precise meaning. This feature enables computer applications to perform sophisticated tasks. A key issue with assigning meaning to the descriptions of web resources is the vocabularies. Many such vocabularies are faceted. Can faceted classification play a role in the Semantic Web? How could SKOS be extended to represent faceted classification? What type of entity is a facet? What are its components? Could faceted classification be formalized to be represented in Semantic Web standards? Could the current KOS evolve to take advantage of the potentialities of Semantic Web technologies? The aim of this paper is to achieve a conceptual model of a faceted classification and its components; to code this model in SKOS and to evaluate such codification. Canonical definitions of “facet” and its components are used to develop a semantic model of a facet schema and its components. Based on this model, a proposal of codification in SKOS is achieved and evaluated.

1.0 Introduction
In her article “A Semantic (Faced) Web?” La Barre asks the question of how to integrate faceted classification into the Semantic Web:

“The chief focus is upon Semantic Web implementations that employ, adapt, or misconstrue the theory or practice of facet analysis and Faceted Classification. A secondary focus is upon suggestions for the creation of operational definitions and functional requirements for facet theory that may serve to enhance, amplify or extend current understandings and practices in Semantic Web implementations.”

(La Barre 2010, 103).

Faceted classification is one of the most important contributions from the knowledge organization to information resources management. Proposed in the twentieth century by pioneers such as Bliss and Ranganathan as an alternative to the rigidity of enumerative classification, faceted classification has gained importance, both academic and practical. Today it is widely used in web architecture, from scientific to commercial sites (Vickery 2008; La Barre 2006, 50), as an information retrieval device (Broughton 2006) and even as a methodology for the development of ontologies (Prieto-Diaz 2002).

Hudon (2019) reminds many authors of the importance of faceted classification applications, such as serving as navigational tool for websites, structuring systems of objects and information about them, and assisting in the understanding of the complex relationships between objects. Broughton and Slavic (2007, 728) stress that “...the potential for faceted approaches to information retrieval in electronic environments had been perceived as early as the beginning of the 1980.”

The web is evolving towards the vision of the Semantic Web, in which resources are not only navigable and understandable by humans, but also by machines. Semantic Web applications navigating between resources can process such resources to perform sophisticated tasks. Within this context, interoperability is key (Zeng 2019), so that generic Semantic Web applications can interact with web resources. Meaning is assign
to such web resources by describing them with different vocabularies. To enable Semantic Web applications to interact with web resources, they must be described formally and with standards languages whose constructs make reference through Internationalized Resource Identifier (IRI)\(^1\) to vocabularies where terms have precise meaning and global scope. Recently several vocabularies developed for information retrieval, library systems or databases have been adapted for Semantic Web technologies and for reference through IRI; many of them incorporate facets. To provide a bridge between Knowledge Organization System (KOS) and the Semantic Web, a standard, Simple Knowledge Organization System (SKOS), a metadata model to describe web resources in the Resource Description Framework (RDF)\(^2\), has been under development; since 2009 SKOS is a W3C standard. However, the present version of SKOS vocabulary does not support the representation of faceted KOS (La Barre 2006, 116) (W3C 2009).

How could SKOS be extended to represent faceted classification? What type of entity is a facet? What are its components? Could faceted classification be formalized in Semantic Web standards (La Barre 2006, 111; Miles and Bechhofer 2008) to take advantage of the potential of such technologies?

The aim of this paper is to represent a faceted schema in SKOS; in order to achieve it, a conceptual model of a faceted classification and its components was developed; such a model was coded in SKOS, and such codification was evaluated.

2.0 The method

Canonical definitions of facets and their components found in KOS literature are used as bases to identify components of a facet. Sources discussing and defining metaphysical entities as classes, subclasses, instances, properties, attributes, characteristics and relationships are used for achieving a semantic model of a faceted schema. Based on this model a proposal for codification in SKOS is developed and evaluated. TemaTres (https://www.vocabularyserver.com/) software is used to generate the codification in SKOS.

3.0 Results

This section contains an analysis of definitions of the concept “facet” to use as bases to develop a conceptual model of a facet schema and its components. On the basis of such a model, a codification in SKOS is achieved.

3.1 What is a facet? What is faceted classification?

La Barre (2003) observed that there is no consensus from scholars for the meaning of term “facet”.

To understanding the meaning of “facet” canonical definitions were selected from KOS literature. While Ranganathan (1967a) emphasizes the aspects of a basic subject and its compound subjects, Mills and Broughton (1977) focus on subclasses and their principles of division, Soergel (1995) sees facets as entities and finally Taylor (1992)

---


2 RDF, The Resource Description Framework, is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data model
discusses a vision of classes specifically concerning their aspects, properties, and characteristics.

Proposed in the twentieth century by Ranganathan and others, faceted classification is an alternative to the rigidity of enumerative bibliographic classifications in which a book is assigned to one class for its general subject. Such classifications are problematic when a book has more than one subject or “points of view” or facets, such as the book “Control of virus diseases of the stem of rice plant in the winter of 1967 in Madras” (Ranganathan 1967b, 13). In faceted classification, each subject within a compound subject is considered a facet in generating by a synthetic process, a notation used to locate the book in a specific shelf and retrieve it. Thus faceted classifications are also information retrieval devices used to retrieve entities’ representations based on their properties—where those are the different subjects of complex subjects as suggested by Ranganathan or product’s characteristics in an e-commerce site.

Ranganathan (1967b, 88) conceived facet as “a generic term used to denote any component - be it a basic subject or an isolate - of a compound Subject, and also its respective ranked forms, terms and numbers.”

Soergel (1995, 258) says, “facets are aspects or viewpoints from which entities - such as food products or subjects (topics, themes) in an area such as education - can be analysed.”

Svenonius (2000, 139) sees facets as categories of generality, defining as “grouping of terms obtained by the first division of a subject discipline into homogeneous or semantically cohesive categories.”

Facets are defined by Taylor (1992, 274) as "clearly defined, mutually exclusive, and collectively exhaustive aspects, properties or characteristics of a class or specific subject" or, by Mills and Broughton (1977, 38), who write "A facet may be defined as the total set of subclasses produced when a class is divided by a single broad principle..." Both definitions introduce the concept of class and its division based on its aspects, properties, or characteristics.

A classification schema is defined as “a list of classes arranged according to a set of pre-defined principles for the purpose of organizing items in a collection or entries in an index, bibliography or catalog into groups based on their similarities and differences to facilitate access and retrieval.” (Fallucchi and De Luca 2018). Another definition of schema classification highlighted by Jacob (2004, 524) agrees that “a classification scheme is a set of mutually exclusive and nonoverlapping classes arranged within a hierarchical structure and reflecting a predetermined ordering of reality.” A faceted schema is also a structure to represent entities and relationships.

According to NISO (2005) faceted analysis is a way of organizing knowledge. Facet analysis is particularly useful for:

- new and emerging fields where there is incomplete domain knowledge or where relationships between the content objects are unknown or poorly defined;
- interdisciplinary areas where there is more than one perspective on how to look at a content object or where combinations of concepts are needed;
- vocabularies where multiple hierarchies are required but can be inadequate due to difficulty in defining their clear boundaries; or
- classifying electronic documents and content objects where location and collocation of materials is not an important issue.” NISO (2005, 13).
Although Ranganathan conceives facets within the scope of the bibliographic classifications of books, contrary to enumerative bibliographic classifications in which to a book is assigned only one subject that defines its position within the classification schema, Ranganathan realized that books could be about several subjects simultaneously. Those component subjects of a compound subject, as previously mentioned, are its facets. A subject is a relationship between a book and what it is about; a component of a subject is a component of any (ontologically) thing a book is about. Conceptual models such as FRBR (IFLA 1997) and LRM (Riva et al 2017) consider subject according to the same ontological view.

3.2 What are the components of a faceted schema?

According to De Grolier (1965, 102), “the term facet itself is just a new, fashionable, word for designating the series of subdivisions of a given subject according to one, and one only, of its characteristics.”

Facet analysis aims to meet users’ needs to access information: “what entities [and] what aspects of those entities are of interest to the user group” throughout the process of conceptual analysis” (Vickery 1960, 11). This is a fundamental criterion in the development of faceted classification.

In an example cited in the ISO 25964-1 (2011, 69) standard, there is a subclass of the agricultural industries products class, dairy products, milk: properties of milk, as milk by fat content, milk by form, milk by source animal; and milk by treatment type. From the milk by fat content property, the following facets are derived, giving rise to subclasses: whole milk, low-fat milk, and skim milk. From the property milk by form are derived the following facets or subclasses: dried milk and liquid milk. From the property milk by source animal the following facets or subclasses are derived: buffalo milk, cow milk, goat milk, and sheep milk; and from the property milk by treatment are derived the following facets or subclasses: condensed milk, evaporated milk, homogenized milk, pasteurized milk, and sterilized milk. To this schema, we added, to help develop our arguments, another class, producer, with three subclasses (two of them milk producers and other a tractor producers), Nestle Massey Ferguson, and Parmalat.

We can distinguish the following elements in this example: the class, milk, one of its subclasses according to a criterion or facet (a milk property, the treatment applied to the milk), milk by treatment; and the subclasses derived from applying this criterion, condensed milk, evaporated milk, homogenized milk, pasteurized milk, and sterilized milk.

Vickery (2008, 156), discussing the structure of a faceted classification, distinguishes elements such as D, a Domain, S1, S2, S3..., Subject fields within a Domain, F1, F2, F3... facets within each Subject field, T1, T2, T3... terms within each facet, and the order of terms within each facet. Within the elements listed by Vickery, Subject fields may be associated with Classes in Taylor’s (1992, 274) previous definition.

Three elements can be identified in these definitions: a class; the set of subclasses, generated based on its aspects; and properties or characteristics.

In the previous examples and in many others, a facet is always defined relative to a class by applying to it a criterion, there is not a facet that is not a facet of a class.
According to Ranganathan (1967b, 55), “Characteristic – an attribute or any attribute complex with reference to which the like or unlikeness of entities can be determined and at least two of them are unlike.”

Facets might be those entities identified in metaphysics and ontological analysis as “properties.” Properties are existentially dependent entities, as their existence depends on the existence of the entities that are the bearers of these properties; a specific marriage cannot exist without the existence of the individuals forming the couple. A specific blue colour cannot exist without being the colour of a specific blue object, such as my blue shoes. Marriage and colour are specific types of properties; the former is a relationship, the last is an attribute (Guarino 1997).

Web languages such as Ontology Web Language (OWL) and the Resource Description Framework Scheme (RDFS) distinguish class-subclass properties, or unary properties, from binary properties. The first type comprises the taxonomic or paradigmatic structure of a domain. The two former types of properties are called in OWL an Object Property (a relationship) and Data Property (an attribute). Object properties require the existence of two entities, one being the domain and the other the range of a specific relationship. A marriage is a typical Object Property. Data Property requires just one entity, the domain of a Data Property, while the range is a set of possible values. A colour is a typical Data Property, as it assigns a data value, “blue” to the property colour of an entity. Sowa (2000, 32) makes same distinction between attributes and characteristics.

“Properties (also called ‘attributes,’ ‘qualities,’ ‘features,’ ‘characteristics,’ ‘types’) are those entities that can be predicated of things or, in other words, attributed to them” (Orilia and Swoyer 2011). To Aristotle (1991, 3) the notion of a property, or characteristic, rests on that of predicates; properties are predicated of subjects; hence they do not exist without being properties of something “(for all colour is in a body” as they are dependent of something to which they are predicated.

The same notion is also in Chen’s Entity-Relationship Model (1976): a domain can be modelled identifying the entities, the relationships and the attributes of entities and relationships.

“A substance—that which is called a substance most strictly, primarily, and most of all—is that which is neither said of a subject nor in a subject, e.g. the individual man or the individual horse. The species in which the things primarily called substances are, are called secondary substances, as also are the genera of these species” (Aristotle 1991, 4).

In this citation, Aristotle relates (first) substance to secondary substances, the genera and the species, defining a class-subclass relationship. A (secondary) substance holds the essence of an entity. Substances are organized in hierarchies of class-subclasses in which subclasses are defined by their essences as having the genus of its parent class plus a differentia from it, in a sequence of increasing specificity. Aristotle (1991, 44) also distinguishes “an accident or property of a thing,” those categories that qualify a subject: quantities, qualities, relations, location, time: “Of things said without any combination, each signifies either substance or quantity or qualification or a relative or where or when or being-in-a-position or having or doing or being-affected” (Aristotle 1991, 3).

Applying a property to a class generates subclasses (classes and subclasses are universals, secondary substances according to Aristotle), the individuals or instances that
make up the extension (Orilia and Swoyer 2011, section 1.1.4) of the class (first substances according to Aristotle).

The property (also a secondary substance) is a subclass of, or is a type of, relates a class and a subclass. The property is an instance of relates a class and its instances. Frické (2010) call this kind of instantiation “first-order instantiation,” or the first-order property. He claims that there is a second kind, the property is an instance of, as in the case of Aristotle citation, the second-order instantiation or the application of second-order properties to first-order properties; that is the case of ‘being a species’ applied to the first order property ‘being a tiger’ – the genera. Tiger is an instance of the second-order type species. According to this perspective, organizing books on poetry, prose, or theatre by literary genre means applying a secondary property (the literary genus) to primary properties (poetry, prose, theatre), the values of literary genus assigned to each book. The concept of second-order property is similar to that of meta-property (Guarino 1997).

Applying Frické conceptualization to the ISO 25964-1 example:
- Shere Khan (first-order instance) -> tiger (genus) -> animal (species);
- condensed milk, etc. (second-order instance, subclass) -> milk by treatment (property) -> milk (class).

A property (or characteristic or facet) divides a class, generating different subclasses, one for each different value of that property existing within the domain. According a facet is a non-sortal or characterizing property of a class (Orilia and Swoyer 2011, section 7.8), generating subclasses but no individuals. Faceted classification is not concerned with the class-subclass hierarchy, a “classification ontology” according to Giunchiglia, Dutta and Maltese (2014, 52) within a domain but rather in a given a class, finding the properties of this class, a “descriptive ontology” of interest to users (Giunchiglia, Dutta, and Maltese 2014, 53), and the instances within the domain with different values to each to these properties. The same notion exists in knowledge organization, between paradigmatic relationships, those permanent, structural or taxonomic relations within a domain, and syntagmatic relationships, those ad hoc, a posteriori or transient relationships (Khoo and Na 2006, 164).

First-order logic and languages such as OWL do not deal well with second order properties as the scope of quantifiers range over individuals (Väänänen 2019). However second order properties are appropriate to specify faceted classification, as they raise not only the individuals that are instances of a class but also the subclasses that are instances of a class, according to Frické (2010).

3.3 How to map components into a conceptual model?

Faceted classification have as components classes, their facets, i.e., their properties (unary properties – subclasses -, and binary properties - relationships), which constitute the criteria to derive instances, and the instances themselves (first- and second-order instances) of each specific facet. Eventually, the order of instances within each facet is also specified. There is also a difference between the two types of facets. One is those derived from classes-subclasses relationships, where the two relata are subclasses of just one primary class within the domain, while the other is those derived from relationships in which the two relata belong, or are subclasses of, different primary classes within the domain.
According to Frické (2010, 44) first-order logic is adequate to discuss second order properties. Applying this conceptualization proposed by the author, plus the OWL concept of properties (Data and Object Properties) of each having a domain and a range. A logic theory is proposed in order to formalize the results of the previous analysis.

- Be \( D_1 \) a domain formed by the primary classes (or unary properties) \( C_1(x_1), C_2(x_2), C_3(x_3) \ldots \);
- Be \( C_1(x_1) \), one such classes;
- Be \( c_1 = \{x_1: C_1(x_1)\} \) (\( c_1 \) is the concept of \( C_1 \));
- Be \( P_1, P_2, P_3 \ldots \) the properties of the class \( C_1(x_1) \);
- Be \( P_1 = \{a_1, a_2, a_3, an \ldots \} \) (\( a_1, a_2, a_3, an \ldots \) are the extension of \( P_1 \), i.e., instances of \( P_1 \)).

- Be \( p_1 = \{x_1,y_1: P_1(x_1,y_1)\} \) (\( p_1 \) is the concept of \( P_1 \));
- Be \( P_1(c_1, p_1) \) i.e., the binary properties with domain \( c_1 \);
- Definition: \( F_1C_1(c_1,P_1) : P_1(c_1, p_1) \) (Facet \( F_1C_1 \) is the class defined by the predicate, or criterion \( P_1 \), having as domain the concept of class \( c_1 \) (\( \{c_1\} \)) and as range the instances of \( P_1 \), i.e. the class defined by instances of the relationship \( P_1 \), as for example, in the case of \( P_1 = \text{milk by treatment} \), the set \( s_1: \{\text{condensed milk, evaporated milk, homogenized milk, pasteurized milk, and sterilized milk} \} \). In this example, as \( \text{milk by treatment} \) is one of the attributes of \( \text{milk} \), \( FC_1 \) maybe a Data Property facet in OWL sense.

- Definition: \( FC_1(c_1,p) : \forall R(c_1,p) \) (the Facets of a class \( C_1 \) are all the binaries properties that have as domain the class \( c_1 \)).

Another example, be \( P_2 \) the property \( \text{milk by producer} \). \( F_2C_2(c_1,P_2) \) is the Facet of \( C_1 \), i.e. the class formed by the set \( s_2 \{\text{Nestle, Parmalat}\} \). In this example, as \( \text{producer} \) is a different class from \( \text{milk} \), \( FC_2 \) may be an Object Property facet in the OWL sense.

- Be \( FC_1 = \{a_1, a_2, a_3, an \ldots \} \) (\( a_1, a_2, a_3, an \ldots \) are the instances of \( P_1 \)).
- Be \( FC_2 = \{b_1, b_2, b_3, bn \ldots \} \) (\( b_1, b_2, b_3, bn \ldots \) are the instances of \( P_2 \)).
- Definition: \( FC_1 \rightarrow \square (an \neq am) \) (two second order instances generated by the same facet cannot be equal).
- Definition: \( F_1C_1, F_2C_2 \rightarrow \lozenge (a = b) \) (it is possible that two second-order instances generated by different facets be equal).

### 3.4 How to map the conceptual model into SKOS?

This section presents the mapping of the conceptual model’s examples of facets in SKOS. Due to page limitations on the number not all the concepts in the example of ISO 25964-1 are presented, only those needed to illustrate the proposal.

```xml
<skos:Concept rdf:about="http://localhost/tematres3.0/vocab/?tema=16">
  <skos:prefLabel xml:lang="en">milk</skos:prefLabel>
  <skos:inScheme rdf:resource="http://localhost/tematres3.0/vocab/"/>
  <skos:broader rdf:resource="http://localhost/tematres3.0/vocab/?tema=12"/>
  <skos:hasfacet rdf:resource="http://localhost/tematres3.0/vocab/?tema=24"/>
  (source_animal)
  <skos:narrower rdf:resource="http://localhost/tematres3.0/vocab/?tema=17"/>
  <skos:narrower rdf:resource="http://localhost/tematres3.0/vocab/?tema=21"/>
  <skos:narrower rdf:resource="http://localhost/tematres3.0/vocab/?tema=35"/>
</skos:Concept>
```

4.0 Concluding remarks

The concept is the fundamental element of the SKOS vocabulary. In basic SKOS, conceptual resources (concepts) can be identified with URIs, labelled with lexical strings in one or more natural languages, documented with various types of note, semantically related to each other in informal hierarchies and association networks and aggregated into concept schemes. Finally, semantic relations play a crucial role for defining concepts, by assigning meaning and context. (SKOS 2009).
Faceted classification was conceived by Ranganathan to deal with subject components of compound subjects; it has been evolving towards handling with different properties of documents – subjects, authors, publication date – and with different properties of things, organizing them in descriptive ontologies. Faceted classification is an information retrieval device. Its effectiveness depends on the analysis and identification of the properties of the objects being described, according to relevance criteria. While taxonomic classification emphasizes the Aristotelian substances, faceted classification emphasizes the accidents, i.e., the different and relevant properties (according to the modelling here proposed) through which the information objects may be retrieved.

The aim of this paper is to develop a conceptual model of a faceted classification and its components; this model is coded in SKOS, and the codification evaluated.

We propose conceptualizing facets as properties (relationships and attributes, attributes and characteristics, data type and object type properties, according to the different authors or sources identified). This modelling option makes explicit the components encompassed by a facet and its interrelations: a class with its properties, a facet (one of the properties of a class, a criterion) and the subclasses or instances created by applying that criterion to the class. Conceptualized in this way, its coding in SKOS follows. The coding maintains and adds to the constructs of the SKOS vocabulary, in addition to using similar constructions as those used in Semantic Web languages, such as RDF, RDFS and OWL, such as classes, subclasses, and properties etc. This helps to bring KOS closer to the mainstream of the Semantic Web.

Acknowledgments:
We are grateful to Prof. Linair Campos for the valuable comments about this paper. This work was carried out with the support of the Brazilian agencies CAPES - Financing Code 001 and CNPq, grant number 305253/2017-4.

References
The Epistemic Communities and Evolution of Knowledge Domains
A Domain Analysis of the Journal Education for Information

Abstract:
Bibliometrics has been presented as one of the approaches to domain analysis. In this context, the relationship between domain analysis and journals has also been explored. There have been several bibliometric studies of specific journals using a domain-analytic approach, often published in ISKO venues, such as those carried out on the journal Knowledge Organization, and the Spanish KO-specific journal Scire. These domain analyses helped to determine the relevance and interest of these journals for the KO and other communities. We propose to study the epistemic communities around the journal Education of Information. An Interdisciplinary Journal of Information Studies. One of the assumptions of our research is that the domain of a given journal is delineated by its epistemic communities. We performed simple bibliometric counts on authors’ affiliations and number of co-authorships in order to identify research elites across the seven periods covering the existence of the journal (1983-2018). We then mapped the co-authorship networks of the journal in order to visualise how its epistemic communities emerged and changed during that time.

1.0 Introduction
The field of knowledge organization is institutionalized, among other things, by professorships at universities around the world, by teaching and research programs at institutions of higher education, by conferences such as the ISKO meetings and scholarly journals (Hjørland 2016). Scholarly journals are of utmost importance not only for the publication and socialisation of research but also for the shaping of the epistemic communities that constitute the domain. Scholars very often do not just consider themselves theoretically and socially integrated into a department or school of a higher institution, but feel more part of the community of researchers that publish, review, and interact in one or several journals that share paradigms and theoretical assumptions underlying their research domains. In this sense, we might say that the authors and the editorial committee of a journal contribute in defining the journal’s domain in perhaps a more influential way than the journal’s scope. Smiraglia (2015, 9) explained the relationship between journals and domains as follows:

“Journals are the formal venues for most scholarly communication, and studying them as whole works is also one means of identifying productive elements of a research front. Of course, few journals are devoted to topical areas that are as narrowly defined as most domains under study. For example [sic], even in the field of knowledge organization, the principle journal Knowledge Organization is devoted to the entire field. Thus, it would likely be the most cited journal in all domains within KO, but there are no journals devoted to specific narrow aspects of KO, such as “integrative levels,” “multilingual thesauri,” or “ethics in KO.””

Of course, this composition is complex and presents almost as many variables related to the definition of the domain as bibliometric indicators can be used for domain analysis.

Bibliometrics was presented by Hjørland (2002; 2017) as one of the approaches to domain analysis. Several authors, including Chen, Ibekwe-SanJuan, and Hou (2010),...
Ibekwe-SanJuan (2008), Ibekwe-SanJuan and SanJuan (2010), and Smiraglia (2015), have explored, in practice, how bibliometrics can be used to map knowledge domains. In particular, Smiraglia has conducted some of the most prominent domain analyses for knowledge organization such as the bookshelf studies of the ISKO international proceedings (Smiraglia 2008; 2011; 2013; 2014; 2017; 2018). Other bibliometric studies have focused on the journal Knowledge Organization (Smiraglia 2012; Guimarães, Martínez-Ávila, and Alves 2015; Alves, Dalessandro, and Bochi 2019) and on the journal Scire (Guimarães, Pinho, and Ferreira 2012; Oliveira et al. 2017). Studies of other journals that also use bibliometric techniques especially in relation to authorship and that could also be considered domain analyses in this vein include those carried out on the journal Scientometrics (Oliveira and Grácio 2012) and Journal of Informetrics (Hilário and Grácio 2018).

Our study aims to identify the epistemic communities formed around the journal “Education of Information. An Interdisciplinary Journal of Information Studies” (EFI) using simple bibliometric techniques that focus on bibliographic units such as authors’ affiliations, co-authorships and research elites. A core assumption of our research is that the domain of a given journal can be delineated by the epistemic communities identified with it. This is especially important for journals that are not highly specialised and have a general scope that might correspond to broad categories such as JCR’s “Information Science & Library Science” or Scopus’s “Library and Information Sciences”; or for journals whose titles suggest an interdisciplinary scope. For instance, the journal under study was called “Education for Information” (EFI) until early 2019 when the subtitle “An Interdisciplinary Journal of Information Studies” was added to it. The word “education” in its title continues to attract submissions from scholars in the field of education despite its aspiration to be an interdisciplinary journal in the broad field of LIS journal as shown by its scope1. This journal welcomes papers on knowledge organization as part of the LIS field and has a good number of ISKO researchers in its editorial board.

2.0 Method

The journal Education for Information was founded in 1983. To perform an analysis of the most productive affiliations, authors, and co-authorship networks, we split its 36 years of existence into seven five-year periods. According to Guimarães, Martinez-Ávila, and Alves (2015), “a five-year period is considered to be an adequate range to characterise scientific production.” Indeed, this was the time range used by these authors to study the epistemic communities of the journal Knowledge Organization. We also believe five-year periods is an adequate range to study the evolution of the 36 years of existence of the journal in seven periods as this number of periods presents a good balance between manageability and a good level of detail for the analysis. Using Price’s Elitism Law (Price 1963) that indicates that the elite of a certain domain (the most productive authors) is represented by the square root of the total amount of authors or publications of the studied domain, we calculated the research elites and the most productive institutions using authors’ affiliations as input.

1 See https://www.iospress.nl/journal/education-for-information/ for its scope.
Price’s Elitism Law has been used in bibliometric studies applied to knowledge organization (Guimarães and Tennis 2012). For the co-authorship networks, we used the software Ucinet version 6.6 and built seven matrices of authors for the seven periods (20x20, 28x28, 57x57, 53x53, 21x21; 49x49; 49x49). We standardised the name of these authors using the authority records of Scopus and the Web of Science when available, in order to avoid redundancies and variations in the names of the same authors.

3.0 Results

The authors in the research elites come from a total of 53 institutions based on their affiliations. Figure 1 shows the nine most productive institutions across the seven 5 year periods. The results show the leading role of the Aberystwyth University, and more specifically the College of Librarianship Wales (CLW) with 29.09% (16) of EFI publications in the first ten years (1983-1992). According to its biographical history\(^2\), the CLW was established in 1964 and it was the first library school in Wales. In 1989, the college merged with the University of Wales, Aberystwyth, becoming the Department of Information and Library Studies. The second institution by order of publications in EFI is McGill University in Canada with 16.37% (9) of the publications distributed in three periods (1983-1987, 1993-1997, 2013-2018). The publications come from three different units at McGill University: the School of Information Studies, the McLennan-Redpath Library, and the Department of Family Medicine. The University of Northumbria at Newcastle in the United Kingdom, the University of Ibadan in Nigeria, Robert Gordon University in the United Kingdom, and Charles Sturt University in Australia, each represented 9.09% (5) of the publications in the seven periods. The contributions of these four institutions come from departments in Library and Information Science. The University of Sheffield in the United Kingdom was present in the first three periods with 7.28% (4) publications while Queen’s University of Belfast and Loughborough University in the United Kindom each accounted for 5.45% (3) of the publications.

Figure 1. Most representative affiliations of the contributions to the journal *Education for Information*

\(^2\) College of Librarianship Wales Archive. Available at: https://archiveshub.jisc.ac.uk/search/archives/cdacfffe-c278-3ac6-aee6-76cf7d465886
Table 1 below shows that in the first three five-year periods (1983-1997), John Andrew Large (henceforth Large JA), who was the first Editor in Chief (EIC) of the journal from 1983 to 2013, was also its most productive author. This points to the prominent role played by the EIC in establishing the journal, expanding and consolidating its authorship networks and thus its initial epistemic community. This assumption is supported by the co-authorship networks we uncovered for the seven five-years periods of the study (see Figure 2 hereafter).

<p>| Table 1. Research elite of the journal <em>Education for Information</em> for the period 1983-2018 |
|----------------------------------|----------------------------------|----------------------------------|
| <strong>Years</strong> | <strong>Authors</strong> | <strong>Articles Published</strong> | <strong>Country</strong> |
| 1983-1987 | Christine J. Armstrong | 7 | United Kingdom |
|           | Blaise Cronin | 5 | United Kingdom |
|           | John Andrew Large | 5 | United Kingdom |
|           | Harold Borko | 4 | United States |
|           | Norman Roberts | 4 | United Kingdom |
|           | Derryan Paul | 4 | United Kingdom |
|           | Gayle Edward Evans | 4 | United States |
|           | Kevin J. McGarry | 4 | United Kingdom |
|           | Richard J. Hartley | 3 | United Kingdom |
|           | Noragh Jones | 3 | United Kingdom |
|           | Marianne Broadbent | 3 | United Kingdom |
|           | M. Wise | 3 | United Kingdom |
|           | John Harris | 3 | United Kingdom |
|           | Michael E.D. Koenig | 3 | United States |
|           | John R. Turner | 3 | United Kingdom |
|           | William Fisher | 3 | United States |
|           | William J. Martin | 3 | United Kingdom |
| 1988-1992 | John Andrew Large | 13 | United Kingdom |
|           | Norman Roberts | 5 | United Kingdom |
|           | Ian M. Johnson | 5 | United Kingdom |
|           | Alan J. Clark | 5 | United Kingdom |
|           | Richard J. Hartley | 4 | United Kingdom |
|           | Joan M. Day | 4 | United Kingdom |
|           | Blaise Cronin | 4 | United Kingdom |
|           | Yves Courrier | 4 | France |
|           | Thomas D. Wilson | 3 | United Kingdom |
|           | Ronald J. Edwards | 3 | United Kingdom |
|           | Kevin J. McGarry | 3 | United Kingdom |
|           | David P. Woodworth | 3 | United Kingdom |
|           | Mary Nassimbeni | 3 | United Kingdom |
|           | Helen Howard | 3 | Canada |
|           | Robin Frederick Guy | 3 | United Kingdom |
| 1993-1997 | John Andrew Large | 8 | Canada |
|           | Thomas A Schröder | 8 | Germany |
|           | Robin Frederick Guy | 7 | United Kingdom |
|           | Thomas D. Wilson | 5 | United Kingdom |
|           | France Bouthillier | 3 | Canada |
|           | Alan J. Clark | 3 | United Kingdom |
|           | Anne Goulding | 3 | United Kingdom |
|           | Marcos Silva | 3 | Canada |
|           | Clive Cochrane | 3 | United Kingdom |
|           | Peter G. Underwood | 3 | United Kingdom |
|           | Douglas Anderson | 3 | United Kingdom |
|           | Flora Smith | 3 | United Kingdom |
| 1998-2002 | Ian M. Johnson | 6 | United Kingdom |
|           | Anne Goulding | 4 | United Kingdom |</p>
<table>
<thead>
<tr>
<th>Name</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Susan Hornby</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>Sajjad ur Rehman</td>
<td>Kuwait</td>
</tr>
<tr>
<td>Charles Oppenheim</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>John Kennedy</td>
<td>Australia</td>
</tr>
<tr>
<td>Rita Marcella</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>Irene Wormell</td>
<td>Denmark</td>
</tr>
<tr>
<td>Graeme Baxter</td>
<td>United Kingdom</td>
</tr>
</tbody>
</table>

### 2003-2007

<table>
<thead>
<tr>
<th>Name</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andrew Kenneth Shenton</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>Mabel K. Minishi-Majanja</td>
<td>South Africa</td>
</tr>
<tr>
<td>Jaya Raju</td>
<td>South Africa</td>
</tr>
<tr>
<td>Ross Harvey</td>
<td>Australia</td>
</tr>
<tr>
<td>Dennis N. Ocholla</td>
<td>South Africa</td>
</tr>
<tr>
<td>Ayoku A. Ojedokun</td>
<td>Botswana</td>
</tr>
<tr>
<td>Sajjad ur Rehman</td>
<td>Kuwait</td>
</tr>
<tr>
<td>Pat Dixon</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>Clive Cochrane</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>Kgomo H. Moahi</td>
<td>Botswana</td>
</tr>
<tr>
<td>Ian M. Johnson</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>John Mills</td>
<td>Australia</td>
</tr>
</tbody>
</table>

### 2008-2012

<table>
<thead>
<tr>
<th>Name</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutawakilu Tiamiyu</td>
<td>Nigeria</td>
</tr>
<tr>
<td>Williams Ezinwa Nwagwu</td>
<td>Nigeria</td>
</tr>
<tr>
<td>Valentini Moniarou-Papaconstantinou</td>
<td>Greece</td>
</tr>
<tr>
<td>Adeola Opesade</td>
<td>Nigeria</td>
</tr>
<tr>
<td>Mary Carroll</td>
<td>Australia</td>
</tr>
<tr>
<td>Afsaneh Hazeri</td>
<td>Iran</td>
</tr>
<tr>
<td>Paul T. Jaeger</td>
<td>United States</td>
</tr>
<tr>
<td>Andrew Kenneth Shenton</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>Kemi Ogunsola</td>
<td>Nigeria</td>
</tr>
<tr>
<td>Bernadette Welch</td>
<td>Australia</td>
</tr>
<tr>
<td>Yun-Ke Chang</td>
<td>Singapore</td>
</tr>
<tr>
<td>Daphne Kyriaki-Manessi</td>
<td>Greece</td>
</tr>
<tr>
<td>Maryam Sarrafzadeh</td>
<td>Iran/Australia</td>
</tr>
<tr>
<td>Emmanouel Garoufallou</td>
<td>Greece</td>
</tr>
<tr>
<td>Naomi V. Hay-Gibson</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>Folake Longe</td>
<td>Nigeria</td>
</tr>
<tr>
<td>Rania Siatri</td>
<td>Greece</td>
</tr>
<tr>
<td>Isola Ajiferuke</td>
<td>Canada</td>
</tr>
<tr>
<td>Georgios A. Giannakopoulos</td>
<td>Greece</td>
</tr>
<tr>
<td>Antonio Muñoz-Cañavate</td>
<td>Spain</td>
</tr>
<tr>
<td>Kanwal Ameen</td>
<td>Pakistan</td>
</tr>
<tr>
<td>Wole Olatokun</td>
<td>Nigeria</td>
</tr>
</tbody>
</table>

### 2013-2018

<table>
<thead>
<tr>
<th>Name</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pierre Pluye</td>
<td>Canada</td>
</tr>
<tr>
<td>Vera Granikov</td>
<td>Canada</td>
</tr>
<tr>
<td>Quan Nha Hong</td>
<td>Canada</td>
</tr>
<tr>
<td>Margaret Max Evans</td>
<td>Canada</td>
</tr>
<tr>
<td>Ya-Fei Yang</td>
<td>Taiwan</td>
</tr>
<tr>
<td>Chih-Kai Chang</td>
<td>Taiwan</td>
</tr>
<tr>
<td>Fatih Oguz</td>
<td>United States</td>
</tr>
<tr>
<td>Kiersten F. Latham</td>
<td>United States</td>
</tr>
<tr>
<td>Bashar K. Hammad</td>
<td>Jordan</td>
</tr>
<tr>
<td>Dennis N. Ocholla</td>
<td>South Africa</td>
</tr>
<tr>
<td>Sharon Stoerger</td>
<td>United States</td>
</tr>
<tr>
<td>Fuziah Mohamad Nadzar</td>
<td>Malaysia</td>
</tr>
<tr>
<td>Isabelle Vedel</td>
<td>Canada</td>
</tr>
<tr>
<td>Jennifer Branch-Mueller</td>
<td>Canada</td>
</tr>
</tbody>
</table>

| Others                        | 1088            |
| Total                          | 1410            |
In Figure 2, red squares represent the authors from the research elites of the journal, blue squares represent the co-authors of the research elite, while the thickness of the lines (edges) represent the intensity of the co-authorships as shown by number of times they co-authored a paper. Scientific collaboration, as a social activity of science, originates from the relations between two or more authors, promoting their visibility, confronting the similarities and differences of their knowledge, and contributing to the emergence of new concepts and domains (Hilário & Grácio, 2011).

In the first period (1983-1987), the network was very sparse and there were few co-authorships, most were made up of no more than two authors and a few isolated nodes (lone authorships). The publications were concentrated among the research elite of the time, notably Large JA, then editor in chief of the journal, and Armstrong CJ who worked at the same institution as Large JA were co-authors in one subgraph. In this first period, the epistemic community of the journal was emerging and not yet structured.

In the second period (1988 to 1992), a more interconnected network became visible with four subgraphs, each of which has three or more nodes of coauthorships and a few lone authors. Although many of the research elite were still in single authorship, multiple collaborations in triads appeared, such as the one between “Edwards RJ, Wilson TD, Roberts N, and Cronin B”. In this period, Large JA co-authored a paper with four authors forming a pentagonal network. This indicates increased collaborations between authors and a structuring of the epistemic communities around the journal.

Price (1963) believed that due to the exponential growth of the number of publications in co-authorship, single authorship would cease to exist. Although single authorship was a significant practice during the first ten years of existence of the journal, we observe that especially after 1993 this practice decreases in favour of co-authorships.

In the third period (1993-1997), the network became even more dense. A huge interconnected subgraph shows Large JA (30) at its center, followed by Smith F (16), Anderson D (13), Underwood PG (12) (see Table 2). These authors were also part of the research elite of this period. Two other people who also held editorship roles in the journal were very centrally placed in the network: Hartley RJ (6) was book reviewer for the journal between 2000-2001 and subsequently joint-editor-in-chief; Guy RF (4) was editor-in-chief together with Large JA at the start of the journal. Two other smaller but dense networks in this period show authors like Wilson TD and Goulding A as central nodes, all of them being part of the research elite.
Figure 2: The seven co-authorship networks of *Education for Information* for the period 1983-2018.
It appears then that in its first three periods of the journal’s existence (1983-1997), the editor-in-chief (Large JA) played a major role in structuring the co-authorship networks and in shaping its epistemic communities.

Bourdieu (1983) stated that scientific capital can be accumulated or even transferred and that it is directly related to the scientific prestige and trajectory of the researcher. This idea reinforces our assumption about the key role of a journal’s editor-in-chief in the emergence, growth and development of a scientific journal.

After 1998, Large JA no longer appeared as part of this journal’s research elite, suggesting that while the journal needed the scientific capital of its founding editor-in-chief to give it prestige during its formative years, once this has been achieved, the journal began to build on this and to consolidate its own scientific capital. This assumption seems to be borne out by the subsequent co-authorship networks after the first three periods (1998-2018). While the journal’s original research elites were predominantly from the United Kingdom (like its first editor-in-chief) and the United States, research elites emerged from other parts of the world (South Africa, Nigeria, Kuwait, Australia, Greece, and Canada).

In the fourth period (1998-2002), the networks continued to display a greater degree of connectivity with most authors being either directly or indirectly connected with the others, with the exception of three small subgroups. The authors with the most co-authorships were Oppenheim C (18) who is placed at the transition point between two interconnected networks. Kennedy J (15), Parker S (15), Hornby S (12) and Morgan S (15) appeared as hubs in the networks.

By contrast with the two preceding periods, the network of co-authorships became very sparse in the fifth period (2003-2007) with a few disparate subgraphs, each with two or three nodes and each disconnected from the others. The journal’s editors who had been present in the preceding periods, and thus instrumental in structuring its epistemic communities were noticeably absent in the network for this period. One wonders why the epistemic communities which had been coalescing around the journal over the first twenty years appear to have disintegrated. This period seems to be a sort of “turning point” (point de rupture) for the journal and may be a sign of disengagement in the journal by its founding editors-in-chief and his network. A more qualitative analysis is required to understand what happened at this time in the journal’s life.

In the sixth period (2008-2012), a new epistemic community seems to emerge comprising two densely connected and moderately sized subgraphs. One completely interconnected subgraph was built around the following authors: Olatokun W, Opesade A, Longe F, Nwagwu WE, Ajiferuke I, Ogunsola K, and Tiamiyu M). Apart from Ajiferioku who is based in Canada, his co-authors are all based in the Africa Regional Centre for Information Science, University of Ibadan, in Nigeria. The rest of the network showed very disconnected subgroups in which the majority of the clusters correspond to the publications that the members of the subgroups shared. Thus, we observe a shifting of the journal’s center of gravity outside the UK and North America.

In the last seventh period (2013-2018), a very dense network appears around Pluye P (20), a professor at the University of McGill (Canada). This dense subgraph is explained by the fact that Pluye P. co-guest edited a series of special issues of the journal on Health information evaluation with a fellow team member Granikov V. Hence, the densely connected subgraph is a direct result of the tradition of multiple co-
authorships within this team from the Department of Family Medicine and the School of Information Studies at McGill University. The other smaller subgroups show fewer number of nodes. The biggest collaborations here have no more than four nodes, as in the case of the members of the research elite Nadzar FM and Branch-Mueller J.

3.0 Conclusion

Grácio (2018) pointed out that scientific collaboration manifests itself through the collective intellectual work, promoting an association of skills and knowledge, uniting researchers with thematic proximity, and sometimes approximating researchers from different areas. Our study showed the crucial role played by a scientific journal’s editor-in-chief in the emergence, development, and consolidation of its epistemic communities which in turn structures research fields. In the case of the journal Education for Information, its first editors-in-chief (Large JA, Hartley RJ) occupied prominent and central positions in the first three periods of the existence of the journal when it needed to gain visibility and credibility as a channel for scholarly publication in the interdisciplinary field of LIS. Once the epistemic communities were consolidated, the influence of its founding editors-in-chief waned and eventually disappeared as they either shifted their attentions elsewhere or left their editorship positions. With the change of editorship in 2014, new more diverse and international epistemic communities began to coalesce around the journal. This indicates that while the journal benefitted from the scientific capital of its founding editors in its first two decades of existence, a point of rupture (split) occurred afterwards in which the epistemic communities built by the founding editors were phased out. As the journal consolidated its reputation, it attracted authors from outside this historic epistemic community, thus arriving at a more diverse and international epistemic communities that are clearly different from those of the founding editors.

References


Chen, Chaomei, Fidelia Ibekwe-SanJuan, and Jianhua Hou. 2010. “The Structure and Dynamics of Co-Citation Clusters: A Multiple-Perspective Co-Citation Analysis.” Journal of the American Society for Information Science & Technology 61, no. 7: 1386–1409.


Knowledge Organisation Systems for Chatbots and Conversational Agents

A Review of Approaches and an Evaluation of Relative Value-Added for the User

Abstract:
Chatbots and smart digital assistants usually rely on some form of knowledge base for providing a source of answers to user queries. These may be more or less structured, ranging from free text, keyword search to interconnected knowledge graphs where answers are surfaced through structured query and/or dynamic logics. This paper will review the state of the art and probe the added value to the end user of using more structured KOSs as the knowledge base component of agent architectures. In addition to the extant literature, experimental development of a chatbot for student information based on a partial knowledge graph is described. A key conclusion is that ontology in a broad sense can be used to drive dialogue as much as to derive factually correct or organisationally-approved responses.

1.0 Introduction
Text and speech-based dialogue agents are growing in popularity as a new or complementary interface to organisational systems and services. They offer high availability, convenience and the potential to identify, triage and satisfy common queries and interactions. As a window into an organisation’s knowledge it seems natural that knowledge organisation systems (KOSs) should play a major role in supporting and directing exchanges, though this is a more sophisticated role than they usually play. For the sake of interoperability and portability, organisations have often erred on the side of simplicity and relatively semantically shallow KOSs to support discovery (Isaac and Baker 2015). However, this level of power is neglecting a wealth of experience in the implementation of Knowledge Representation and Knowledge Engineering coming from the computer science / AI disciplines (Giunchiglia, Dutta, and Maltese 2014). These somewhat deeper approaches allow the formal expression of conceptual structures and their interrelation, as well as the application of automated logical inference over a system’s knowledge base to arrive at evidence-based conclusions. These affordances once again become potentially useful when it comes to conversational agents, so it is a good time to revisit standards and tools developed for knowledge representation and apply them to chatbots (Cameron et al. 2018)

A similar driver has been the developments in natural language interfaces, where some progress has been made in mapping unstructured queries to formal data relations. Indeed, there was a call to action at ISKO 2004 for better connection between KO and computational linguistics (Jones, Cunliffe, and Tudhope 2004). John’s et al.’s exploratory work showed how term disambiguation and the mapping of user queries to KO structures could be improved by using a general purpose thesaurus and expanding matching to include thesaurus scope notes.

The application of KOS to conversational interfaces requires an active approach and a pragmatic view of semantics (as the concepts are action-oriented and application-specific) in addition to a unification of user and organisational warrant (Hjørland 2007), as
the goal is to allow the user to interact successfully in real-time with an organisation’s knowledge structures.

This paper will look at ways in which KOS can both fuel and drive dialogue in conversational agents and chatbots. This will be based on an exploration of ways in which the user experience at the dialogue interface is affected by bot behaviours and capabilities. We will then look at examples of how KOSs have been used in domain-specific dialogue systems, before exploring in more detail a case-study on a bot for student information. It is hoped that these sections will serve to highlight challenges and opportunities in KOS-driven agents. Firstly, however, we will look at a broad typology of bot technical architectures to help illustrate where KOS traditionally sits.

2.0 Dialogue System Typologies and Components

A common distinction made with agents are goal-driven or rules-based and those that are more open-domain and able to tackle a wider variety of topics. (Altinok 2018). The former are perhaps more likely to end in satisfactory completion due to having a more predictable, constrained structure that can also be driven in part by menus and fixed choices. They are however restricted in requiring explicit representation of the dialog structure (Augello et al. 2013) and in only being able to perform a limited set of functions – also, if the pattern matching fails, the bot may flounder. In contrast, in some of the more open AI-enabled systems, the most suitable response is predicted based on models trained on a large dataset of human dialogue. Here, some precision is often sacrificed in an attempt to provide the user with a more open ended, flexible conversation experience.

In most cases, where a dialogue requires an informational response, then a data or knowledge base is central to the architecture (Figure 1), whether this is based on the endpoint of a retrieval-based dialogue fragment or the model-guided prediction of an AI-based agent.

Increasingly, bot architectures include natural language processing to elucidate the “things” or entities being talked about and the goals or intents of the user in relation to them. These enriched structures are used to (perhaps fuzzily) match to entries in the knowledge base using similarity or information retrieval algorithms. The best matches can then be used to derive the system response.

In recent years we have also seen an increase in more hybrid architectures, including both rules-based and generative components and it is likely that this trend will increase, as the two approaches can to some extent make up for the shortcomings of the other. Certainly, new dynamic architectures are needed to help lift chatbots out of the “trough of disillusionment” in which many still lie – as the user experience at the interface falls short of expectations.
3.0 User Interaction with Conversational Interfaces

In terms of interaction patterns, Hill, Ford and Farreras (2015) looked at interaction logs from conversations with chatbots in comparison to human partners, and found that chatbot interactions contained fewer words per message but a greater number of messages overall compared to humans. This indicates that a system needs to be able to proactively seek clarifications, in the same way that humans do in everyday interactions (Schlangen 2004). These may be to enable the completion of semantic underspecification, or to clarify intent.

A related issue is dialogue initiative. Finite state (rule-based) systems tend to dominate the initiative, whereas in real conversation, initiative alternates between participants (Jurafsky 2018). One-sided initiative feels inflexible, less natural and can’t handle statements containing multiple pieces of information.

Reasons for using a spoken agent include time saving and multi-tasking (Luger and Sellen 2016). Luger & Sellen noted that regular users invest time in understanding the agent’s strengths and capabilities, including initial playful interactions. Interestingly, factors affecting interaction included having sufficient understanding of an agent’s capabilities and internal workings, including “whether or not its capabilities altered over time” and the extent to which the agent was learning from the user.

Maintenance of context is a key fundamental strength of human conversation, and an automated agent’s limited abilities in this area can be a source of frustration (Vtyurina et al. 2017). The “anaphora” or maintaining the implicit referent of the conversation continues to be a technical challenge for agent developers.

Vtyurina et al.’s study also highlighted the need for information credibility to be strengthened in an agent’s answers, perhaps by providing links where information could be verified, similarly where a variety of opinions exist on a topic, users asked for summaries of these.
4.0 Evaluation criteria for Conversation-Oriented KOS

Several dimensions have been proposed to organise KOS themselves, one of the more common spectra being semantic strength – this itself correlated with time and money (Souza, Tudhope, and Almeida 2012). At the higher semantic strength end, formal theories and ontologies are heavier and provide richer expressivity, logics and entailments or automated reasoning. Conversely with weaker semantics comes the opportunity for more interoperability at the syntactic level.

Ontologies themselves can vary according to domain or application specificity and in the area of chatbots this maps quite well onto generalist versus specialist conversational capabilities.

Soergel (2001) proposes a set of evaluation characteristics for KOS, including purpose, coverage, conceptual structure, extent of precombination, access and display, and degree of updating.

In terms of coverage, Tennis (2013) notes the temporal limitations of correctness and that dynamic theories are necessary given the inevitability of change and the need to cater for the ever changing and evolving knowledge domain. A geologic or archaeological view might best reflect how KOs change over time in practice, though the degree to which change is incremental or immediate may vary between systems and stakeholders. In the perhaps more radical change scenario, interaction with users may be seen to drive immediate change.

In sum, many existing criteria for KOS evaluation focus purely on aspects of the KOS itself rather than its wider position within an organisational information system. This neglects pragmatic aspects such as: ease and automation of updating, degree of fit with legacy corporate systems, degree of redundancy. These latter considerations are surely essential if the KOS is to be sustainable and to fulfil a central role in corporate digital communications.

5.0 KOS Approaches for Agents

Perhaps the simplest form of data store to support an agent is question-answer pairs, where the user query is compared to the knowledge base to find the closest matching question. This kind of architecture was used as a core component, for example, by Carisi et al (Carisi, Albarelli, and Luccio 2019) to build an airport information chatbot, though their bot also was able to connect to other airport databases to report flight information. Such a knowledge base can be bootstrapped using FAQ lists from web pages or from call centre records.

More “capable” bots may be closely linked to underlying ontology. Athreya et al’s dbpedia bot (2018) used third party natural language parsing to match queries to DBpedia URIs and literals, with literals given precedence as more likely to correspond to factual answers.

Altinok (2018) describes a system for banking services that uses an ontology, both to represent the bank’s services and their features, but also to structure the dialogue. Noun phrases in user queries are scanned for known entities by matching to ontology classes and individuals (with use of a synonym list to capture alternative phrasing). Matched entities are then used to provide the context for further queries, somewhat mitigating the anaphora problem. Conversation memory is also used to keep track of ontology nodes that have been visited in order to avoid repetition and provide the next dialog prompt.
For their bot to discuss issues in climate change, Groza and Toniuc (2017) used an ontology but converted it to API.AI pattern matching model for use in bot interaction. Classes were mapped to entities and roles to intents. However, when no intent was matched (i.e. queries were not easily matched to models), entailment rules were used derived from a debate corpus.

Roca et al. (2019) propose a general microservice architecture for their prototype chatbot to support patients with chronic illness. The argument is that such bots are easily extensible without disruption to service. In terms of communication and storage, the bot used AIML to represent the dialogue logic, with FHIR (Fast Healthcare Interoperability Records) as the data store for the patient and associated data entities (e.g. Care Plan, Practitioner, Organization, Questionnaire), which could be used to generate monitoring questions – via dynamic AIML generation - and personalise the conversation.

6.0 Case Study: Student Personal Circumstances Procedure

We have developed a prototype chatbot for students, based around common queries connected to timetabling, choice of optional modules, changing course and so on. In doing so, we have worked closely with our enquiry services to understand both the kinds of enquiry received and how students are subsequently advised in accordance with university regulations.

One of the most common sources of enquiry is around problems taking exams or submitting coursework due to illness, bereavement or other life events. These cases may be eligible for special consideration in accordance with the university’s marks removal or personal circumstances procedures depending on whether or not the student has submitted the work or attempted the exam. To be eligible, students need some evidence to support the claim, or they may self-certify once per academic year without supporting evidence. Examples of the type of queries around these processes are shown in Table 1.

<table>
<thead>
<tr>
<th>Source</th>
<th>Example</th>
<th>Appropriate Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information Point Advisors</td>
<td>“I want to self-certify but where is the form?”</td>
<td>From QA knowledge base</td>
</tr>
<tr>
<td></td>
<td>“I don’t have any evidence, what should I do?”</td>
<td>Triage Self-Certification</td>
</tr>
<tr>
<td></td>
<td>“When is the deadline to submit a PC form?”</td>
<td>Triage Personal Circumstances</td>
</tr>
<tr>
<td></td>
<td>“Can I make a request after the deadline?”</td>
<td>Clarification of Intent</td>
</tr>
<tr>
<td></td>
<td>“I sat an exam but I shouldn’t have, what can I do?”</td>
<td>Triage Marks Removal</td>
</tr>
<tr>
<td></td>
<td>“I handed in the work but I shouldn’t have, what can I do?”</td>
<td>Triage Marks Removal</td>
</tr>
<tr>
<td></td>
<td>“How do I submit a request?”</td>
<td>Clarification of Intent</td>
</tr>
<tr>
<td></td>
<td>“I applied for missed assessment but I would like to cancel it now, what should I do?”</td>
<td>Triage Process Cancelation</td>
</tr>
<tr>
<td>Web Search Logs</td>
<td>“personal circumstances evidence submission”</td>
<td>Clarification of Intent</td>
</tr>
<tr>
<td></td>
<td>“personal circumstances but can’t attend campus”</td>
<td>Triage Personal Circumstances</td>
</tr>
</tbody>
</table>

Table 1: Sample queries around personal circumstances supplied by information point advisors or taken from web search logs (in some cases these have been amended slightly to preserve anonymity). System appropriate response noted.
Our initial prototype used a rule-based decision tree for personal circumstances to guide the user through the process, once the applicability had been established through identifying that intent (using a set of intent patterns such as “sick exam” or “missed deadline”). The logical dialogue structure for this is shown in Figure 2. Additionally, a “one-off” question-answering dataset is used when there is a strong match to a more specific question relating to the topic (here the user “passes through” to a direct answer rather than entering a protracted dialogue.

Problems with the approach taken include that, while the structure correctly formalises the regulations, a) the structure is hard-coded and any changes to the process require reworking at the code level; and b) it follows a set script or funnel down which all users are sent. It therefore does not adequately accommodate short-cut, clarification and process-related questions such as “I have submitted by work and don’t have any evidence, what should I do?”, “What counts as evidence?” or “What to do if I don’t have evidence?”, or “How many times can I self-certify?”. While problem a) was addressed by allowing new dialogue structures to be loaded from external files at run-time, the essential inflexibility of this approach remained.
We therefore have commenced work on a second prototype using a domain ontology to represent the enquirer (student), the processes we want to represent, and the requirements for these processes. The dialogue can then be built based around the completeness of the entity (object properties) as constructed during the conversation, a more “frame-based” approach. If these are missing, then this will trigger information gathering questions. Once enough information is available, then the enquirer can be channelled into the application procedure for the relevant process.

In order to facilitate maintenance, we store the natural language options to identify intents in the entities RDF labels, and the queries to generate in the entity and property RDF comments. We also take advantage of SWRL rules which capture the formal regulations (e.g. Listing 1). Here, the student is flagged as eligible for marks removal if they have: a) submitted already b) have admissible evidence.

Listing 1: Example SWRL rule for Marks Removal Eligibility

\[
\text{Student}(s) \land \text{Assessment}(a) \land \text{Process}(p) \land \text{Documentation}(d) \\
\quad \land \text{hasSubmitted}(s, a) \land \text{hasDocumentation}(s, d) \land \text{isAdmissibleForProcess}(d, p) \\
\rightarrow \text{isEligibleForProcess}(s, \text{MarksRemoval})
\]
7.0 Discussion

Our case study is typical of many agent interaction settings as: a) the initial user input may be minimal and requiring clarification and b) it is very important the user is advised appropriately once the system is sufficiently appraised of its circumstances. It therefore needs to match user status to organisational policies and processes – a strong argument for use of a formal, rule-based ontology as a core architectural component. That said, for it to fulfil the requirements of a conversational interface, it does need to be sufficiently detailed, pointing toward the expensive, resource-intensive end of the KOS spectrum (Souza et al. 2012).

Considering Soergel’s (2001) evaluation criteria for KOS in the light of agent applications, we see that the purpose requires relatively high specificity and that, in terms of sources and coverage, while preferred terms can be those defined and used by the organisation, they should be comprehensively mapped to the natural language of the user in order to be correctly identified, matched and prioritised.

In terms of conceptual structure, while some depth of class hierarchy seems useful for organisational purposes, it seems less likely that these are very useful for satisfying user needs, as queries that are applicable to ancestor classes are probably more unusual. Instead, the properties of a specific atomic entity seem to be more useful to guide slot filling and frame completion.

Finally, and perhaps most importantly is Soergel’s criterion “updating”, which we should elaborate to include not only frequency but ease of updating, along with the centrality of the KOS itself to the overall organisation information system. It is here that ontology-based agents seem to be very early-stage – very little is said about how they can be maintained or embedded within an organisation, though semantic wikis offer a useful opportunity to bridge the gap between knowledge engineering and domain expertise. On the ontology creation side, there is a relatively large amount of work looking at how they can be derived from textual sources, but it is almost inevitable that further quality assurance steps will be needed and that this cannot be fully automated. Architecturally speaking, Roca et al.’s example (2019) seems to point the way, in describing an approach that allows for the knowledge base to be incrementally enhanced without loss of service. In their case, however, underlying knowledge structures are relatively partitioned or siloed, which would seem to risk increasing the problem of maintainability. On the flip side, use of single integrated ontologies will require version control and continuous validation to ensure that extensions to not introduce errors in existing rules.

8.0 Conclusions

This paper has looked at KOSs within chatbot and agent architectures in order to understand what the KOS itself can contribute to the end user experience. Clearly, there are demonstrable ways that a KOS-driven service can not only represent the domain knowledge with which a user is interacting but can also shape and facilitate the dialogue itself. Here there are potential advantages in goal-driven or regulatory oriented applications in maintaining formal correctness, where “generative” AI-based systems cannot necessarily guarantee such accuracy.

While individual prototypes and technical approaches are relatively well documented, less has been published on how to get to a state where suitable ontologies are embedded within an organisation and maintained as a core system. While perhaps less
superficially exciting, this work can establish a basis from which conversational interfaces can evolve productively.

References


Call Us by Our Name(s)
Shifting Representations of the Transgender Community in Classificatory Practice

Abstract:
The linguistic shifts in terms used for self-representation and collective representation by the transgender community are a compelling exemplification of subject-based classification schemes’ shortcomings. These shortcomings become readily apparent in digital environments which enable user-generated tagging of content; on these platforms, members of the transgender community are able to maintain their preferred self-referential terms to establish the subject(s) of their work. In using creator-generated tags as a source for contemporary terminology to be used in subject-based classification schemes, I will build upon previous work by surveying self-referential terminology generated by creators of original works about the transgender community. By comparing the terminology used in hashtags on posts centered on the transgender community on Instagram with the terminology present in the Library of Congress’ subclass heading for HQ77.7—HQ77.95 and the Library of Congress’ available subject headings relating to gender variance, I will examine the shifts in rhetorical representation of materials focused on the transgender community over time, using this as justification for additional documentation regarding global changes made to classification schemes and indexing languages into online catalog records.

1.0 Introduction
The Library of Congress Classification Scheme is the bibliographic classification scheme developed by the Library of Congress. Its primary sites of implementation are libraries within the United States of America, and it is also implemented in numerous libraries within and outside the United States (Olson 2000, 54). The flaws built into LCC have been well-documented throughout the 20th century—particularly as these flaws relate to the scheme’s representation of marginalized communities (Adler 2017, 9-10, 44; Fox 2016, 375; Olson 2000, 54; Olson 2001, 541). The transgender community is one of many groups that has been represented within LCC by terminology which is perceived by the community as outdated and pejorative (Roberto 2011, 57). LCC subclass HQ77.7—HQ77.95 is where materials about the transgender community are classified using the term “transsexualism,” which is considered to be an outdated—and often offensive—term by many transgender individuals.

The terms used to name LCC classes and subclasses are meant to reflect the language used by relevant discourse communities; however, despite terminological shifts within queer studies and other fields, which have been catalyzed by the transgender community’s increased political and cultural visibility, “transsexualism” continues to be a term used within LCC. I argue that the HQ77.7—HQ77.95 subclass heading is no longer reflective of the terminology used by transgender individuals during acts of self-representation, using hashtags on Instagram as a way to measure the prevalence of particular terminology within the LGBTQ community when creating and sharing content by and about transgender individuals. I also argue for the addition of version control in public-facing catalog records as a means for documenting large-scale edits made to knowledge organization infrastructures, so as to increase transparency and facilitate institutional accountability in classificatory and cataloguing practices.
2.0 Literature Review

When creating a bibliographic classification scheme—particularly those which use a hierarchical structure—one must establish different concepts’ relations to each other (Tennis 2012, 1351), and also the terminology used to name concepts and subjects. It is understood that language is always imperfect and constantly evolving, and this is true in classificatory practices as it is in other contexts (Ranganathan 1937, 62; Tennis 2016, 574). Despite this acknowledged fluidity, which is a natural result of knowledge production, authors of classification schemes are put in a position where they must choose stable terms according to common usage (Adler 2017, 34; Roberto 2016, 63). In the implementation of a classification scheme, there must also be a way for terms to systematically be updated to reflect changes in common usage (Fox 2016, 380; Olson 2001, 660-661; Ranganathan 1937, 64). Changes to classification schemes and indexing languages are unavoidable as knowledge is constantly produced, and these changes may be based upon literary warrant and/or ethical reasons (Tennis 2013, 2). This process of addition and revision broadly prevents the catalog from becoming obsolete (Ranganathan 1937, 66). Further, periodic revisions to classification schemes and indexing languages can and should prevent the continued use of terminology which is harmful to marginalized communities (Roberto 2016, 63).

Literary warrant is a key concept in consideration of how terms are chosen in classification schemes; it is the idea that relevant literature should be surveyed with attention given to the language used to describe phenomena within a given knowledge domain, so that the terminology reflects the established discourse within that domain (Johnson 2010, 662). Literary warrant may be a useful way of choosing a particular term, as it is intended to enforce the language used within a community of experts (Adler 2009, 313; 2017, 20; Olson 2000, 56); however, the enforcement of literary warrant in practice often undermines its conceptual usefulness, as inconsistent application of this concept may also justify the use of outdated or derogatory terminology (Olson 2000, 57-58, 65).

The HQ section of the Library of Congress classification scheme is a space which has been subjected to scrutiny by a number of scholars and researchers within the field of knowledge organization (Adler 2015, 489; Johnson 2010, 663). Materials about the LGBTQ community—as well as about sexual deviations, sadism, masochism, and fetishism—are catalogued in the HQ70s, with materials about “transsexualism” residing in HQ77.7—77.95 (Library of Congress, n.d.). According to the application of literary warrant within this subclass of LCC, “transsexualism” is a valid term based upon literary warrant from early 20th century medical and psychiatric literature, in which the term appears frequently when referring to gender-variant individuals (Adler 2009, 668; 2017, 36; Johnson 2010, 668). However, contemporary discourse communities use other terminology to describe individuals who do not identify and/or present as their birth-assigned gender; most notably, these discourse communities include the transgender community, who are shaping the ways in which gender variance is named by creating and disseminating their works in both analog and digital media.

‘Transgender’ as a broad term used to describe individuals who transgress traditionally-codified notions of gender may be attributed to Leslie Feinberg, a prominent transgender author and activist (Adler 2009, 318-319); the term was initially introduced by Virginia Price in the 1980s to describe individuals who do not present as their birth-assigned gender yet do not seek to medically transition (Johnson 2010, 666). Today,
“transgender” is a preferred term to “transsexual” by most members of the transgender community; this preference is validated in common speech as well as in textual works in many disciplines and authoritative entities. While “transgender people” was introduced into LCSH in 2007 (Library of Congress 2019a), “transsexualism” remains the term used in LCC to describe this community.

3.0 Method

Past work which focuses on the terminology used to describe materials about the transgender community has engaged with terminological disparity between LCC/LCSH and discursive practices within the LGBTQ community. Melissa Adler’s study of the disparities between LCC/LCSH and user-generated tags in platforms which are intended for use with book recommendations is a notable example of research focused on how the transgender community is named in different formal and informal environments; the users of these platforms are readers and reviewers of the books which are given tags to facilitate retrieval (Adler 2009, 320). Adler’s study shows the disparities between terminology used in LCC/LCSH and terminology used by readers who engage with materials by and about the transgender community.

To build upon Adler’s work, I have focused on terminology used by the creators of materials about the transgender community on Instagram. As a social media platform, Instagram is used primarily for disseminating images or short-form text. To facilitate content retrieval, creators affix hashtags to their post; there is no controlled vocabulary imposed upon hashtags, meaning that creators may use whatever terms they wish in hashtags, but creators also have an interest in applying hashtags which are common-usage terms if their intention is for their post to be seen by a wide range of individuals. Instagram was chosen for this work because of the importance that social media platforms and other digital environments have for the LGBTQ community; these digital environments are relied upon by the LGBTQ community for sharing not only creative works, but relevant information pertaining to legislation, politics, and medical care.

Instagram enables a user to search for a term as a hashtag, and these searches return all of the posts upon which this term is affixed as a hashtag as well as the total number of posts on the platform which bear the hashtag. Beginning on October 8, 2019, eleven terms were entered as searches for hashtags. These eleven terms were chosen based upon the terminology used in LCC/LCSH as well as upon the derivations of Library of Congress terminology which are used upon this platform. “Transsexualism” is the term used to name HQ77.7—77.95, and “transgenderpeople” and “transsexuals” are terms used in LCSH (Library of Congress n.d.; 2017; 2019a). “Transgender” and “trans” were chosen based upon common usage among the transgender community in published works as well as in spoken discourse; “transsexual” was chosen to maintain continuity in studied terms’ forms. To account for terminological disparity within LCSH, which has subject headings for “female-to-male transsexuals,” “male-to-female transsexuals,” “transgender men,” and “transgender women,” (Library of Congress 2018a; 2018b; 2019b; 2019c) qualifications regarding the gender binary were affixed to “trans,” “transgender,” and “transsexual” as search terms. The respective quantities of posts returned on October 8, 2019, with each of these hashtags were recorded as the initial quantities; the same eleven terms were searched as hashtags on February 5, 2020, so as to assess the relative frequency of use of each respective term.
4.0 Findings

The number of posts bearing “trans” or “transgender” as hashtags at the beginning of the data collection process was at least double the quantity of the number of posts bearing “transsexual” or some derivation thereof; this was seen to be the case regardless of whether or not these terms were qualified by some directional acknowledgement of the gender binary. Further, the frequency at which posts were tagged with “trans” or “transgender” was higher than that of posts tagged with “transsexual.” This was seen not only in all three of these terms in their singular and plural forms, but also in derivations of these terms which included directional acknowledgement of the gender binary.

<table>
<thead>
<tr>
<th>Searched Terms</th>
<th>Number of Posts, Oct. 8, 2019</th>
<th>Number of Posts, Feb. 5, 2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>femaletomaletrans</td>
<td>5,442</td>
<td>5,788</td>
</tr>
<tr>
<td>femaletomaletransgender</td>
<td>18,159</td>
<td>19,043</td>
</tr>
<tr>
<td>femaletomaletranssexual</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>maledofemaletrans</td>
<td>3,332</td>
<td>3,565</td>
</tr>
<tr>
<td>maledofemaletransgender</td>
<td>6,810</td>
<td>7,849</td>
</tr>
<tr>
<td>maledofemaletranssexual</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>trans</td>
<td>7,053,350</td>
<td>7,441,758</td>
</tr>
<tr>
<td>transgender</td>
<td>8,965,528</td>
<td>9,315,772</td>
</tr>
<tr>
<td>transgenderpeople</td>
<td>1,975</td>
<td>1,999</td>
</tr>
<tr>
<td>transsexual</td>
<td>593,165</td>
<td>605,412</td>
</tr>
<tr>
<td>transsexualism</td>
<td>367</td>
<td>372</td>
</tr>
<tr>
<td>transsexuals</td>
<td>28,942</td>
<td>30,267</td>
</tr>
</tbody>
</table>

Figure 1: Eleven terms searched as hashtags on Instagram, accompanied by each term’s respective quantities of tagged posts.

It is particularly notable that, over the course of data collection, the only terms which showed no increase in posts using them as hashtags were “femaletomaletranssexual,” and “maledofemaletranssexual,” and that only five posts were tagged with “transsexualism” over the course of seventeen weeks. Considering that these terms were chosen due to their presence in LCSH as subject heading terms and as a subclass heading in LCC, this should be particularly noteworthy.

5.0 Discussion

As seen in the collected data, “transsexualism” and derivations thereof is a term which is used far less frequently than “trans” or “transgender” to signify posts about gender variance on a popular social media platform for purposes of retrieval by platform users. This extension of literary warrant to include self-representative discursive practices in digital environments calls into question the continued use of “transsexualism” as the term used to name HQ77.7—77.95.

A bibliographic classification scheme’s implementation necessarily affects the ways in which patrons access materials and interpret materials (Adler 2017, 5; Tennis 2012, 1355). By maintaining “transsexualism” as a subclass heading, despite the ongoing work
of the transgender community, the broader LGBTQ community, and activist librarians to shift LCC/LCSH terminology, these organizational infrastructures are enabling a particular interpretation of works about the transgender community (Tennis 2012, 1355). This interpretation prioritizes early 20th century medical and psychiatric discourse over the transgender community’s self-representational discourse, with the effect of marginalizing the latter (Adler 2017, 34; Olson 2000; 55, 68-9). Further, this maintenance of an outdated, non-preferred term as a subclass heading may be detrimental to patrons who are using preferred terminology when attempting to access materials about this community (Roberto 2011, 63). Based upon the discursive practices of the transgender community, the Library of Congress must address their reliance on what the majority of the described community perceives to be an inaccurate term.

The relative speed of terminological shift from ‘transsexual’ to ‘transgender’ as a preferred term for the transgender community is made manifest in LCC, in which a once-common term’s presence is now perceived as a problem. This is a particularly interesting example of term shift, as ‘transsexual’ and ‘transsexualism’ are terms which have a history of being used not only by doctors, psychiatrists, and researchers, but also self-referentially by many gender variant individuals in printed works and spoken language alike. When one considers the pace at which the transgender community began to adopt the term ‘transgender’ self-referentially; the fact that ‘transsexual’ is not universally perceived as a pejorative term by gender variant individuals, as the collected data shows that it is still a used term (albeit with less frequency than “trans” or “transgender”); and the increasing use of the term “trans”, a truncation of transsexual and transgender, within the community; the implication is that this community may continue to engage in ongoing self-referential terminological shift. LCC’s inability to adequately express this shift can—and should—be understood as a vestige of antiquated notions of gender and gender variance (Roberto 2011, 57). This example also can be understood as an example of the problems which may be caused by a universal classification scheme which is so firmly fixed that it cannot accommodate for new knowledge or conceptual shifts (Olson 2000, 57; Tennis 2016, 574).

The findings in this study are consistent with those of other knowledge organization researchers who have lent their academic focus to terminological stagnancy and contentious terms’ presence in LCC/LCSH (Drabinski 2013, 98; Fox 2016, 375; Olson 2000, 74; 2001, 647), particularly with regards to terminology used to describe the LGBTQ community (Adler 2015, 491; 2017, 324; Roberto 2011, 63). Future research concerning the continued use of obsolete and demeaning terminology within LCC and other classification schemes must continue, as the findings of researchers often bolster the lived experiences of individuals outside of academia and may act as justification for broader structural changes to be made to LCC and other classification schemes. This future research must not focus solely on justification for updating specific terminology. While changing a given term may be a welcome short-term solution, updating terms on a one-to-one basis will not solve a host of broader concerns that this particular study illuminates.
6.0 Reducing Opacity in LCC/LCSH

At this time, any modifications made to catalog records which represent global changes to LCC/LCSH are not made visible to patrons, who only see the finalized catalog record. A valuable addition to public-facing library catalog records would be additional documentation regarding global changes to LCC/LCSH as they affect these records. Rather than making changes relating to terminology and erasing all traces of former terminology, it would be useful to maintain a visible history of these changes on the records themselves, so as to convey the history of modifications made to this classification scheme and indexing language. The addition of version control would also be copacetic with LCSH itself, which includes subject headings’ edit history in its entries.

LCC and LCSH are information infrastructures which are fixed in the context during which they were created and modified; they are also representations of this institution’s perspective on concepts, topics, and subjects at that point in time. This is unavoidable, as it is not possible for cataloguers to ever truly be neutral (Drabinski 2013, 95; Olson 2000, 64; Roberto 2011, 62), nor is it possible for cataloguers to predict the future directions of knowledge production, self-representational practices, and so on. Version controlling public-facing library catalog records would provide support for navigating the retrieval-related issues caused by terminological shift, as this would convey both outdated and updated terminology while prioritizing the latter (Tennis 2012, 574). Identities and the terms used to express them are multifaceted and constantly changing—both at an individual level and at collective levels—and knowledge is constantly being produced, yet LCC/LCSH maintain their fixity without acknowledging the temporal nature of the materials over which they exert control (Tennis 2012; Olson 2000, 57; Roberto 2011, 60). If the ways in which these changes affect classification schemes and indexing languages were documented and made accessible, users would still be able to rely upon familiar (if outdated) terminology while being encouraged to adopt terminology which is more current.

When considering how to reduce the harm perpetuated by LCC/LCSH’s antiquated notions of gender identity and sexual orientation, Emily Drabinski advocates for the maintenance of these infrastructures, along with increased interactions in which librarians use these existing infrastructures as a way to teach patrons about institutional control over non-normative ways of being (Drabinski 2013, 108-109). The assumption made is that patrons will notice the workings of LCC/LCSH and ask librarians about them, and that librarians will be equipped to answer their questions, which cannot be guaranteed on any scale; more importantly, this suggestion enables the continued use of infrastructures which have been revealed to be harmful in numerous capacities. By documenting large-scale changes to LCC/LCSH in item records, Drabinski’s idea of using LCC/LCSH as tools for studying institutional representations of identities and communities would still be possible, but in a way that still encourages large-scale modifications of these infrastructures. These traces of oppressive and harmful practices would be usable as a tool for having instructional conversations about how institutions exact power in what are currently ‘invisible’ ways without maintaining these infrastructures in ways that perpetuate harm.

Numerous scholars have called for increased transparency in the acknowledgement of terminological shifts and other changes made to classification schemes and indexing
language (Olson 2000, 68; Tennis 2013, 6; 2012, 1352; Roberto 2011, 63). Adding public-facing version control to item records in a library catalog would, quite literally, convey these global changes to users in this and other contexts. The assumption here, though, is that there will be global changes to document; put another way, the usefulness of version control in library catalogs’ item records is contingent on the Library of Congress’ willingness to make global changes justified by literary warrant and/or demanded by the communities being named.

At this time, LCC/LCSH are exerting control over how users seek and retrieve materials, but in a way which is largely invisible to the average user (Drabinski 2013, 97; Olson 2000, 66). Increased transparency about how these infrastructures are exerting that control—in ways which are both beneficial and detrimental—would be a way for users to better understand how these infrastructures are shaping their efforts to find information. The addition of this documentation to public-facing library catalog records would thus be a mechanism for institutional accountability, as version controlled records would indicate reparative practices and harmful practices alike. LCC/LCSH have been widely criticized for their damaging treatment of the LGBTQ community and other already-marginalized communities by researchers and librarians over the course of the 20th century; transparency about the continuation of harmful practices may encourage the institution to actually implement the reparative changes which have been repeatedly demanded. That being said, changes have been made to LCC/LCSH which do reduce the harm caused by problematic terminology and other practices (Olson 2000, 60), and version controlled item records would convey these positive changes to users as well.

The well-intentioned goal of LCC/LCSH are to facilitate browsing and retrieval of relevant materials, but the frequency with which these infrastructures are shown to fall short of this goal necessitates a closer look at the tools used for this purpose (Tennis 2016, 578). These tools include the terms used to name classes and subjects, but also the catalogs themselves. Future research may focus on the logistics of exploiting digital environments’ capabilities for enabling the addition of version control to library catalog records on a global scale, so as to reduce the burden that additional documentation could place upon cataloguers, librarians, and other relevant information professionals.

The addition of version control to library catalogs’ public-facing item records would surely be beneficial to the transgender community and other communities which have been poorly represented by LCC/LCSH, but would also be beneficial to any subject represented within a universal classification scheme; knowledge is constantly changing within disciplines, and any work which seeks to increase a classification scheme’s flexibility and ability to accommodate these changes is sure to benefit users across disciplines. That being said, the addition of version control to catalog records will not, by itself, solve problems such as terminological stagnancy and continued use of derogatory language. This documentation will convey this institution’s shifts in perspective on the bibliographic universe and enable transparency and accountability on an institutional level; it will provide support to users within and outside academia who are using these catalogs, and will hopefully facilitate future work which will substantially reduce the maintenance of oppressive practices in information infrastructures.
7.0 Conclusion

The evolution of self-representative terminology used by the transgender community, which may be seen in not only published works but in discursive practices in digital environments, has not been adequately reflected in the Library of Congress Classification Scheme, which still uses the outdated term “transsexualism” to name the subclass within which works about this community are classified; this community is only slightly better represented in Library of Congress Subject Headings. This disparity in terminology conveys previously-voiced concerns about the continued use of outdated, perjorative terminology in LCC/LCSH; it also shows the greater shortcomings of knowledge organization infrastructures which are so fixed that they cannot accommodate for terminological shift based upon the development of new knowledge. The addition of public-facing documentation of global changes to these infrastructures to catalog records may ameliorate these concerns, as this documentation would convey changes in representations of subjects; it would also act as a mechanism for institutional accountability to its patrons, particularly those who have called for changes to be made to these infrastructures.

References


A Critique of the Use and Abuse of Typologies in Cultural Policy Analysis

Abstract:
Typologies used explicitly or implicitly in cultural policy analysis (CPA) do not comport well with basic principles of typology. The research area of CPA therefore suffers from a lack of conceptual rigour. In order to provide an understanding of why CPA finds itself in this unfortunate state, this article examines three specific and relevant typologies, and concludes that CPA would benefit from using principles of knowledge organisation more actively.

1.0 Introduction
The typologies examined in this article provide the framework for a significant proportion of the research conducted within the social science subdiscipline of cultural policy analysis (CPA). More precisely, these typologies are utilised in various analytical attempts to address fundamental CPA research questions such as: How does the public sector legitimise public funding of the arts? As a research area, CPA is predominantly empirical and critical, and to a lesser degree conceptual. There is thus a potential for increasing the conceptual rigour of CPA by applying principles of knowledge organisation.

As Hjørland (2017) has pointed out, it is possible to make a long list of units that can be classified. Although objectives are not on Hjørland’s list of potential units of classification, we may add them to this list. For example, researchers within education might find it valuable to classify educational objectives (Bloom 1956). Correspondingly, researchers within CPA classify cultural policy objectives and cultural policy rationales. Thus, CPA includes both the construction of typologies and the process of assigning empirical observation to them.

Researchers within CPA rarely develop their typologies themselves. More often, they use typologies a priori, for example in addressing questions about the ways in which authorities legitimise public funding of the arts. A priori typologies are applied in analyses of empirical material such as policy documents and interview transcriptions. However, the typologies that are applied a priori to answer these kinds of research questions are rarely scrutinised explicitly by the researchers who make use of them. This lack of explicit scrutiny might be due to the tendency within the social sciences to take

---

1 In the relevant literature, “cultural policy studies” is a more common term than “cultural policy analysis”. However, the term “studies” in “cultural policy studies” is often a signal that the publication is intended to be critical, as in critical theory (e.g. McGuigan 2004). “Cultural policy analysis” is a more general term and includes analyses of cultural policy that are informed both by critical theory and by other theoretical perspectives. In this article, which considers cultural policy analysis broadly understood, the term “cultural policy analysis” is therefore more accurate and useful than “cultural policy studies”. However, the abbreviation CPA (cultural policy analysis) is neither an established nor a commonly used abbreviation. It is used in this article for the sake of brevity.

2 This can be observed in prominent CPA journals, such as the International Journal of Cultural Policy, which only rarely publishes articles with the stated aim of scrutinising analytical frameworks.
typologies for granted. As Bailey (1994, 83) has remarked, “because classification is so ubiquitous, it is relatively easy to overlook it”. The methodological apparatus of CPA is derived from several social science disciplines, but most notably economics and sociology. These two disciplines differ considerably in terms of their approaches to concepts. The typologies that these two disciplines offer will be discussed here with regard to two basic requirements of typologies, which are that they should consist of collectively exhaustive and mutually exclusive types (Marradi 1990) and that they should be practically useful.

2.0 Typologies applied in CPA

We will consider three typologies that are frequently applied by researchers within CPA. The conceptual rigour of these typologies is important for the general rigour of CPA, owing to the prominence of these typologies within it.

2.1 Welfare economics

A seminal work within CPA was published in 1966, when Baumol and Bowen (1966) provided an analysis of the economic challenges of the performing arts. The authors stated that the purpose of their study was to “be able to specify objectively the alternatives facing the arts and to describe their costs and the burdens they require society to shoulder” (Baumol and Bowen 1966, 4). This publication marked the founding of cultural economics as a research area. Accordingly, Baumol and Bowen’s book is a starting point for many researchers who wish to conduct CPA from a welfare economics perspective. Their goal was to “describe the logic on which such a decision, one way or the other [public funding or no public funding of performing arts], should be based if it is to satisfy the criterion of rationality” (Baumol and Bowen 1966, 378).

The logic that they referred to was the logic of welfare economics. Based on this logic, they presented a typology of three possible arguments for public support of the arts:

1. Egalitarian grounds (“public funds devoted to the opening of opportunities to the impecunious”; see Baumol and Bowen 1966, 379).
2. The education of minors (“The arts must be made available early, while tastes are still being formed and behavior patterns developed”; see Baumol and Bowen 1966, 380).
3. The public goods argument of non-excludable and non-rivalrous goods (Government subsidies of the production of a good are warranted if it is both impossible to exclude consumers from using the good and additional consumers do not reduce quality of consumption for other users; see Baumol and Bowen 1966, 381).
Now, how does this typology fit as an *a priori* tool for analysing and comparing ways in which governments legitimise public support of the arts? The short answer is: not so well. An immediate problem with the “logic” of this typology is that it is easy to suggest additional potential political arguments for public support of the arts that are plausible and yet are not covered by the typology. In other words, the types in the typology are not collectively exhaustive. For example, several authors have pointed out that, historically, public support of the arts has been significantly legitimised through the argument that national pride is increased by excellent arts performances (e.g. Bennett 1995, Duelund 2009). Some Danes, for example, might get an increased feeling of national pride after witnessing an excellent performance by a Danish orchestra of a symphony written by a Danish composer. But how does that argument fit in Baumol and Bowen’s (1966) overview? The “national pride” argument cannot reasonably be claimed to have been covered by the typology. Another example of a potential argument that is not covered is the frequently made claim that the economic output of a nation or region can be increased by governments “investing” in the arts for the sake of economic impacts (e.g. Myerscough 1988).

Contemporary cultural economics has developed somewhat from the overview in Baumol’s and Bowen’s seminal book. For example, Fullerton (1991) lists rationales of public funding that are related to concepts such as *redistribution*, *merit goods*, *public goods* and *externalities*. Publications within contemporary welfare economics (e.g. Stiglitz and Rosengard 2015) frequently assume, although implicitly, that these kinds of legitimations are collectively exhaustive and mutually exclusive. In that regard, contemporary welfare economics represents a more developed *a priori* tool for CPA than that of Baumol and Bowen’s (1966) typology. But, while the typology of public funding rationales provided by recent welfare economics might be satisfactory with regard to collective exhaustiveness and mutual exclusiveness, welfare economics is still not a *practically useful* tool in policy analysis. The typologies of legitimations for public
funding that are provided by welfare economics are rarely applied as *a priori* analytical tools in policy analysis. The reason is probably a lack of phenomenological relevance. For example, if we consider Fullerton’s list referred to above, an analyst using welfare economics as an *a priori* tool might ask herself: Is the social cohesion that a hypothetical government claims to provide through public funding of the arts an example of a “public goods” argument or of an “externalities” argument for public funding? The problem in answering this is that the technical and abstract way in which the different types of welfare economics arguments for public funding of the arts are labelled significantly reduces their practical applicability within CPA. The lack of practical applicability of welfare economics is due to the fact that ministries of culture, along with other funding bodies, rarely formulate themselves in ways that can easily be interpreted through the lenses of economic concepts such as “public goods” and “externalities”.

In conclusion, when it comes to providing CPA with an *a priori* tool, what seems to be the major strength of economics as a discipline is that the over decades it has improved Baumol and Baumol’s typology into a typology that consists of mutually exclusive and collectively exhaustive categories of legitimation (e.g. Stiglitz and Rosengard 2015). However, a major weakness of welfare economics, from the perspective of CPA, is that the labels that it attaches to different types of legitimation of public funding of the arts are too technical and abstract. The labels lack ‘thickness’ (Geertz 2017). Hence, it is difficult for CPA researchers to assign empirical observations to the types of legitimation that welfare economics provides.

### 2.2 The intrinsic–instrumental dichotomy

A second typology that has been extensively applied within CPA is the *intrinsic–instrumental dichotomy* (e.g. O-Kyung 2010). Although this is not always explicitly stated, this dichotomy derives from classical sociological theory – most prominently, Weber’s typology of types of social action. Weber’s typology of social action consists of instrumentally rational, value-rational, affectual and traditional action (Weber 1978). Researchers within CPA have frequently made use of the first two of these types, differentiating between *instrumental* value and *intrinsic* value as motivations behind cultural political argumentation, where these two notions correspond to the first two of Weber’s categories.

The dichotomy is frequently used as an *a priori* analytical tool in CPA, but the dichotomy is also contested within CPA, for two reasons. First, within CPA publications, the notion of instrumentality has *de facto* become inextricably associated with value-laden CPA, in that the term “*instrumental* policies” has become synonymous with “*undesirable* policies” (e.g. Belfiore 2015). This means that it is has become difficult to use the term “*instrumental*” in a value-free manner (see Weber 1949). Second, the intrinsic–instrumental dichotomy is conceptually contested, in the sense that researchers within CPA disagree on whether it is at all plausible, in any context, for a political body to make the claim that a policy is motivated by the intrinsic value of art. Some researchers have concluded that it is logically fallacious to apply the proclaimed intrinsic value of art as an argument for public funding (e.g. Bakhshi *et al.* 2009; Culyer 1973;
Vestheim 2007). Now, let us assume that a researcher would like to conduct a value-free form of CPA, and to ask a research question based on the intrinsic–instrumental dichotomy, such as: To what extent is contemporary cultural policy (for example, in Denmark) underpinned by the assumption that cultural policy is instrumental; and, by contrast, to what degree is contemporary cultural policy underpinned by the idea that the arts have intrinsic value? Going further, the CPA researcher will now have to answer a number of non-trivial questions: What does the researcher understand by the term “intrinsic value”? What does the researcher believe that the authorities mean when they convey their intentions to legitimise public funding of art through the intrinsic-value argument? Does the researcher assume that the term “intrinsic”, whether this is communicated literally or intentionally, refers to a value that is independent of the existence of human beings? Does the notion of the “intrinsic value” of art invoke religion or spirituality?

In other words, the intrinsic-value concept elicits a plethora of follow-up questions. This is because the concept seems to be notoriously elusive, in that researchers who rely on its relevance rarely provide an intensional definition. What seems to be the case is that the concept is either defined in the negative, as the opposite of instrumental value, or through a form of “family resemblance” (Wittgenstein 1958) or extension (Marradi 2012). Researchers seem to assume that readers will know what intrinsic value is all about, although the readers themselves cannot be assumed to be able to define it by its intension. For example, one researcher concludes that the arts are “intrinsically valuable to society” (Belfiore 2004, 200, emphasis in original), without defining intrinsic value. Without imputing too much to this quote, the assumption seems to be that readers do not need a definition of intrinsic value because they will know from practical experience what intrinsic value is. But if researchers wish to analyse the degree to which cultural policy is instrumental – and, by contrast, the degree to which it is underpinned by notions about the intrinsic value of art – it is reasonable to expect a conceptual elucidation of the concept of intrinsic value. Hence, applications of the intrinsic–instrumental dichotomy within CPA are likely to provoke more questions than they answer.
What this tells us is that if a researcher within CPA states, explicitly or implicitly, that he or she will apply the intrinsic–instrumental dichotomy as an *a priori* analytical tool, this is a decision that requires at least two follow-up questions: (1) Is the use of the intrinsic–instrumental dichotomy in a CPA publication intended to convey that the publication is value-laden, or is the dichotomy used in an analytical, value-free manner? (2) If the researcher’s aim is to analyse the intentions in the empirical material, how does the researcher define the term “intrinsic value” in terms of its intension? The problem with publications within CPA is that these questions are rarely asked, and thus rarely answered. Hence, the intrinsic–instrumental dichotomy is a rather confusing *a priori* tool that lacks conceptual rigour.

### 2.3 The orders of worth

The final *a priori* CPA tool that deserves to be scrutinised here is the so-called *orders of worth* framework provided by Boltanski and Thévenot (2006). This framework is much more recent than the two previously discussed tools, but its prevalence within CPA during the last decade or so has been considerable.

Boltanski and Thévenot state: “we have been able to observe the operation of six higher common principles to which, in France today, people resort to most often in order to finalise an agreement or pursue a contention” (Boltanski and Thévenot 2006, 71, emphasis in original). Subsequently, they describe how their observed “common principles” extend into various “political forms of worth” (Boltanski and Thévenot 2006, 83–124). In other words, the common principles are considered to apply to the political realm, not just to the social world in general. It is this application of the common principles to the political realm that has paved the way for the entrance of the “orders of worth” framework into CPA as an *a priori* tool.

From the perspective of CPA, there are problems with this system that have not been sufficiently acknowledged by researchers. According to Boltanski and Thévenot themselves, “our list of principles is not exhaustive; we can discern the shape of other polities that might be constructed” (Boltanski and Thévenot 2006, 71). In other words,
if we apply the “orders of worth” framework as an *a priori* tool in the analysis, we cannot simply assume that all possible justifications of public funding of the arts are covered by Boltanski and Thévenot’s (2006) system. The system is not intended to be an exhaustive list of the various forms of legitimation that are used in relation to public policies. In addition, it seems clear that the types of justifications that are identified in their book are not mutually exclusive. The authors describe six different polities, or “worlds”. These are the “inspired world”, the “domestic world”, the “world of fame”, the “civic world”, the “market world” and the “industrial world”. However, nowhere do Boltanski and Thévenot claim that these worlds are supposed to be mutually exclusive. On the contrary, it seems reasonable to conclude that they are not. For example, there is nothing in Boltanski and Thévenot’s definitions that suggests that agents cannot be associated with both the “world of inspiration” and the “industrial world” simultaneously. This also means that the different ways of justifying an action, such as the action of public art funding, cannot easily be assigned to the typology of worlds from which these justifications emanate, because one specific form of justification might emanate from more than one world.

Let us look at some studies that use the “orders of worth” system in an *a priori* manner. One study concludes that “the world of inspiration is the most important value regime in the creative industries” (Nijzink et al. 2017, 609). Presumably, artists and creative workers are therefore of the opinion that the “world of inspiration” should inform the public authorities that fund the arts. But this is trivial. It is not necessary to have an advanced education in the social sciences to predict that, of Boltanski and Thévenot’s various “worlds”, artists will prefer the “world of inspiration”. In addition, Nijzink et al. seem to misunderstand Boltanski and Thévenot’s conceptual system by implicitly referring to the system as a typology of collectively exhaustive and mutually exclusive forms of legitimation. They simply rank the different worlds in terms of their importance as cultural political legitimations for different shareholders (Nijzink et al. 2017, 609), without much regard for the openness of the system that is emphasised by Boltanski and Thévenot. In another study, Lemasson concludes that “the coexistence of the inspired logic with the civic or the industrial ones” has been “difficult to achieve” in cultural policy in Quebec (Lemasson 2017, 81). But this is just a way of concluding, in a language coloured by the framework of Boltanski and Thévenot, that the arts sometimes need public support to survive. Yet another study concludes that, of Boltanski and Thévenot’s orders of worth, it is the “civic world” that has dominated the legitimating rhetoric of cultural funding in Norway and Sweden at the beginning of the 21st century (Larsen 2016, 129). But does this mean that “the inspired world” has become less important? Where are the borders between the “inspired world” and the “civic world”? Has artistic excellence become less important, while inclusion of different demographic groups of society has become more important? If the latter is the case, then how does it bring us deeper insights to embed these observations in Boltanski and Thévenot’s conceptual system? This seems entirely unclear.

In summary, Boltanski and Thévenot’s orders of worth appears, at first glance, as a productive *a priori* tool in analysis, owing to the authors’ use of the phrase “on justification”, for example in the title of their book. The phrase “on justification” might suggest that the system is tailor-made for analysing how public bodies justify taxpayer funding of their services. However, as an *a priori* tool in CPA, this system has two major
weaknesses. First, the system is not intended to consist of collectively exhaustive and mutually exclusive types of legitimations. Second, the system is made up of labels that tend to obscure empirical findings more than they provide insights into them. Thus, the main benefit of the orders of worth system, from the perspective of the CPA publications that make use of it, is not conceptual rigour. The system has managed to “conquer the academic scene” in CPA (Mangset 2010, 48), but it seems reasonable to ask whether there have been guru effects involved in the conquest (Sperber 2010, Elster 2011).

3.0 Conclusions and implications

This article has questioned the ways typologies are used in an a priori fashion within a specific area of the social sciences, namely cultural policy analysis (CPA). The aim has not been to provide an exhaustive presentation of typologies that are applied by contemporary researchers, nor has the goal been to provide an exhaustive examination of the methodological problems associated with the typological systems examined. Instead, the goal has been to show that Bailey’s (1994) warning – namely, that we should not take our typologies for granted – needs to be continuously repeated in the social sciences.

The problem addressed in this article might be due to a too deep divide between theoretical and empirical investigations. It is important to avoid a situation where theory within the field is used as “sacred texts to be worshipped as totems” (Turner 1998, 245). The construction of typologies is ongoing work, and better typologies of cultural policy rationales can be mined out in a collaborative effort by the research community as a whole. Thus, analysts should be encouraged to reflect on the typologies that they use, and to do so explicitly in the methodological sections of their publications. It might be tempting, especially for younger researchers, to adopt a readymade typology, in particular, if they observe that this typology has already reached paradigmatic status. But such a paradigmatic status should not be allowed to deflect researchers within CPA from explicitly scrutinising their typologies.

References


Juan Bernardo Montoya-Mogollón – São Paulo State University – UNESP, Brazil
Sonia Troitiño – São Paulo State University – UNESP, Brazil

Digital Forensics Science and Knowledge Organization: An Interdisciplinary Approach to Addressing the Conceptual Challenges of Born-Digital Records

Abstract:
The digital age has drastically impacted the way people think and act using technology. The advancements of the 20th and 21st centuries have ushered in a number of changes related to the treatment of information. Private and public organizations have faced increasing pressure to shift to digital formats. However, the speed at which these changes have occurred has created several conceptual concerns related to the trustworthiness of digital records and systems, and the security afforded by their infrastructure. This paper identifies and addresses several of these concerns through the interdisciplinary theoretical frame of Archival, Diplomatic, and Digital Forensics areas. This alternative approach is presented as a contribution to Knowledge Organization (KO), which can help reframe the way scholars understand the following problems posed by digital records: 1) The challenge of maintaining digital records as authentic, accurate, and reliable sources of information; 2) The challenge of establishing trustworthy repositories for information storage and access; and 3) The challenge of ensuring long-term maintenance and preservation of records. We argue that the different concepts and techniques of Archival Diplomatics and Digital Forensics, such as provenance, original order principles, and chain of custody, have much to contribute to the professional study and practice of KO in digital environments, particularly where it relates to issues of authenticity.

1.0 Introduction
Currently in Archival Science, a series of terms are used as synonymous with ‘records’ in bureaucratic and archival scenarios, such as data, document, and information. Some of these terms were created by The International Research on Permanent Authentic Records project – InterPARES. We decided to adopt them as they can be used in an interoperable kind in Archival-Diplomatics, Forensics Science, and Knowledge Organization (KO).

We will need to use this terminology in this article because in several studies these words are written without distinction, which can cause ambiguity for our research problem. We will look at a series of disciplines that use these concepts in their own field, sometimes with a different meaning.

Data is the smallest meaningful piece of information. A document is recorded information. A record is any document created (i.e., made or received and set aside – i.e. kept, saved – for action or reference) by a physical or juridical person in the course of practical activity as an instrument and by-product of said activity (in some countries, a record is called archival document or simply document). Information is a message intended for communication across space and time (Duranti and Thibodeau 2006, 15).

1 This group born in 1990 with the “aims at developing the knowledge essential to the long-term preservation of authentic records created and/or maintained in digital form and providing the basis for the standards, policies, strategies, plans of action capable of ensuring the longevity of such material and the ability of its users to trust its authenticity.” (www.interpares.org)
Consequently, we will be using the concept of born-digital records. A search of academic literature can find electronic records or electronic documents, but a small difference exists: electronic regularly makes reference to electric signals and is applied when we talk of hardware or technological devices (Prayudi and Ashari 2015, 1). While the meaning of ‘digital’ is considered as the interaction of various hardware and software components to give rise to binary digits: “Hardware and low-level software detects the physical properties and interprets them as binary digits – i.e., digits that can take only one of two possible values – which are called bits. By convention, we say that the two possible binary values are 1 and 0” (Lee 2012a, 511). Therefore, we will use the word ‘born-digital’ to define those records created/produced exclusively in digital environments and not as a part of a digitization process.

Even today, almost forty years after the introduction of digital records in business systems, there are still problems considering these born-digital records as legally reliable. Archival Science is concerned with studying and proposing measures to be adopted for their maintenance and long-term preservation. Maintenance and preservation were an issue with the analogical or paper record, but a machine did not need to mediate the process as is the case with digital records. Therefore, Archival Science is constantly renewing, by searching knowledge in other areas to understand the nature and composition of born-digital records and making efforts to maintain their trustworthiness.

For this purpose, we intend to establish the relationship between digital Archival-Diplomatics, Digital Forensics Science, and KO. This contribution is contemplating how Digital Forensics represents the data storage in several technological systems and provides a knowledge framework to understand the composition of born-digital records. With this knowledge, the archivist will have the ability to preserve and access records, in addition to possessing other kinds of tools found in digital forensics.

2.0 Archival-Diplomatics Science

Archival Science deals with three core concepts: provenance, original order, and chain of custody:

“Two principles key to archival theory are the principle of provenance, which respects the documentary facts, maintaining the records of one creator separately from another, and the principle of original order,
which mandates keeping/describing the records in the order in which they were created and used. When these principles are respected and articulated through archival description, the authenticity of the record aggregations is protected [18-20]. The presumption of authenticity derives initially from the context of creation and chain of custody, and the documented processes of establishing intellectual, administrative, and usually, physical control – appraisal, accessioning and archival arrangement” (Rogers 2013, 8).

As Guimarães and Tognoli (2015, 567) also note, both provenance and original order concepts that belong to “respect des fonds”, are relevant to Archival Knowledge Organization (AKO). These concepts allow the identification of the 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” (Tognoli and Rodriguez 2018, 46-47). Thus, the identification process is key to the creation of knowledge as it must consider the application of the principles of provenance and original order.

On the other hand, Diplomatics Science was founded in 1681 by Jean Mabillon (1602-1707), a French Benedictine monk who established the foundations to analyze the authenticity of diplomas (hence the name Diplomatics for the records of this period) kept in the Abbey of Saint-Germain-des-Prés. He wrote his masterpiece de re diplomatica libri VI, at a time of social upheaval due to document forgeries throughout Europe. His work aimed to examine and compare the extrinsic (support, ink, seal, signature, etc.) and intrinsic (the intellectual message of the author) elements of the form of diplomas or records to verify their authenticity or falsehood.

In the 20th century, Diplomatics Science was renewed and adapted for Archival Science. The two scholars key to reviving the study of the relationship between Diplomatics and Archival Science were Hilary Jenkinson (1957) and Roger-Henry Bautier (1961). However, in the early 1980s with the arrival of the digital age, the extrinsic and intrinsic elements of form began to address the concept of authenticity in the digital world (Duranti 1989; Duranti 2009, 42). This involved analyzing the individuals that participate in the process of record creation, the contexts in which the record exist, the act or transaction in which it participates, the procedures and documentary forms governing its creation, and the relationships that connect it to other records (Rogers 2015, 8).

However, in the digital environment, the extrinsic and intrinsic elements of form are not easy to address because the representation of the digital records operates in abstract structures of conceptual, logical, and physical layers. The conceptual layer is the way the record is understood by a person on the screen or monitor. The logical layer is an object that is recognized and processed by hardware and software. The physical layer is an inscription of signs on a physical medium. The conceptual layer is closer to the traditional record in paper format. The logical layer is the most important to understand how the record is represented because of the storage data (content) and the metadata which could be easily visible in the conceptual layer or by using specialized tools (Rogers 2015).

Archival-Diplomatics has been broadly studied in KO by a series of academics. We intend to continue this trend by considering the application of forensics science as a contribution to the KO. The study of Digital Forensics Science has been significant to analyze and preserve the authenticity of born-digital records in the long-term.
3.0 Digital Forensics Science

Digital Forensics is a relatively new discipline (Pollitt 2010), which arose at the start of the 1980s at the height of computer accessibility. At the same time, illicit acts started to appear. Currently, this discipline is faced with an increasing number of challenges because of the internet and the complexity of digital devices. The most widely used definition of Digital Forensics is provided by the Digital Forensic Research Conference -DFRWS:

“The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations” (Palmer 2001, 16).

In its early days, it was known as Computer Forensics, but with the broadening of its research objective, its name was changed to Digital Forensics: digital because it analyzes the information configured in binaries digits (as explained earlier) and forensics because it has a narrow relationship with the law. The etymological meaning of forensics is related to the forum “because the courts are forums where information may persuade us to restrict or remove individual liberties, they have proven to be a serious testing ground for scientific research” (Palmer 2001, 2).

The relationship (and necessity) between Digital Forensics and Digital Archival Diplomats draws on maintaining and preserving the trustworthiness of born-digital records. In a study written by Elizabeth Diamond (1994, 140), she considered that “if the historian is the lawyer in the court of history, then the archivist is the forensic scientist”. The responsibility of archivists is to ensure the maintenance and preservation of records, more-so, of born-digital records. A series of skills are required, for example: 1) Knowing the internal framework or layers; 2) How it is represented in the technological systems; 3) How to maintain its trustworthiness over time; and 4) How to provide access to users.

According to Adam Jansen and Luciana Duranti (2011), knowledge of professionals about digital records must be more accurate since the challenges in the digital environment are increasingly complex:

“Custodians can only preserve records as trustworthy (i.e. reliable, accurate and authentic) as they are when first created. It is therefore the custodian’s responsibility to establish the identity of the records prior to acquiring them and to maintain that identity, together with their integrity, afterwards (MacNeil, 2004). In the digital environment, this is a tall order, because it is not possible to preserve digital records; it is only possible to preserve the ability to reproduce them (Duranti and Thibodeau, 2006). As it will always be necessary to retrieve the binary bits and process those bits through the use of intermediaries (i.e. hardware and software) in order to render the evidence into a human readable format, it falls upon the custodian to ensure that the necessary intermediaries will exist when needed. To render representations with an accuracy that is able to withstand a diplomatic analysis requires the custodian to store the binary content of the record, including indicators of all the elements of documentary form necessary to convey the essence of the record, in a manner that ensures the record will be rendered with the same presentation and in the same context that gave it meaning” (Duranti and Jansen 2011).

The above quotation explains the responsibility and accountability of archivists. Therefore, Archival-Diplomats provides a framework that supports the preservation of the integrity and identity of born-digital records. The junction with digital forensics allows the use of tools and practices to deepen the representation of digital devices and its layers by keeping accurate information to produce accurate knowledge.
4.0 Digital Forensics Science and Knowledge Organization

We noticed an evolution in how Archival Diplomats fit the digital environment. The efforts made to work interdisciplinarily and apply the knowledge of disciplines, such as Diplomats, Forensics, Law, History, and Computational Science, have supplied prolific answers for issues related to the trustworthiness, maintenance, and preservation of digital records in determined social contexts. The application of Digital Forensics to Archival Diplomats processes allows us to interchange concepts, techniques, and tools to understand the structure of digital environments. Therefore, it should be stated that the archivist and the professional in Digital Forensics process have similar objectives for different purposes. The author Corinne Rogers (2013, 6) notes this similarity between the identification of (digital) records in the archival study and the (digital) evidence in the study of digital forensics. She expresses that:

“Archivists and digital forensics practitioners share challenges involved in appraising and analyzing large volumes of digital material. The core archival functions have been identified as appraisal and acquisition, arrangement and description, retention and preservation, management and administration, reference and access [11]. Digital preservation has been demonstrated to encompass records creation and recordkeeping [12], thereby extending the archival functions over the entire life cycle of digital records. The traditional archival functions may be compared with the functions of digital forensics practice: identification, preservation, collection, examination, analysis, presentation and decision [13]. At the root of each is investigative research into the material in question – namely the digital traces of activities, and the relationships of those traces to the actors and actions which gave rise to them” (Rogers 2013, 6).

The functions mentioned above in both Archival-Diplomats and Digital Forensics are shared and complemented by concepts such as provenance, original order, and chain of custody in the digital context. However, their meanings are slightly different. Provenance is defined as the identification, extraction, and the saving of essential information about the context of creation; original order reflects original folder structures, files associations, related applications, and user accounts; and the chain of custody is the documentation of how records were acquired, whether or not they were transformed, and the use of hardware and software mechanisms to ensure that the data has not been inadvertently changed. Another relevant concept shared between forensics and archival sciences is the identification of sensitive information, specifically personal and private information, “the same tools that are used to expose sensitive information can be used to identify, flag and redact or restrict access to it” (Lee 2012b, 5).

Therefore, the chain of custody is paramount in the digital environment. This element is essential to maintaining the integrity of the bitstreams, and although some challenges exist to ensure such integrity, its use is increasingly common. Its issues are related to 1) The volatility of digital evidence in resources such as register, memory, table, processor, temporary filesystem, disk, remote logging and data monitoring, physical configuration, and network topology, as well as archived data (Prayudi and Azhari 2015, 3); 2) with medium failure/bit rot; and 3) the obsolescence of both software and hardware (Lee et al. 2012).

Digital Forensics practitioners build knowledge by deepening the internal structure of the born-digital records with the aim of better understanding its nature. To get to this point, it is necessary to understand the inner workings of logical, physical, and conceptual layers addressed by Rogers (2015). In addition to this, we need to examine how these layers are represented and structured within the technological systems. The author Christopher Lee (2012a) outlines the structure of these layers in an interesting overview.
He defines nine levels of representation (Figure 2) described as digital resources, namely: 0) bitstream on physical medium; 1) raw signal stream through I/O equipment; 2) bitstream through I/O equipment; 3) sub-file data structure; 4) file as “raw” bitstream; 5) File through filesystem; 6) In-application rendering; 7) Object or package; and 8) Aggregation of objects.

Figure 2: Digital Resources – Levels of Representation (Lee 2012c).

<table>
<thead>
<tr>
<th>Level</th>
<th>Label</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>Aggregation of objects</td>
<td>Set of objects that form an aggregation that is meaningful encountered as an entity</td>
</tr>
<tr>
<td>7</td>
<td>Object or package</td>
<td>Object composed of multiple files, each of which could also be encountered as individual files</td>
</tr>
<tr>
<td>6</td>
<td>In-application rendering</td>
<td>As rendered and encountered within a specific application</td>
</tr>
<tr>
<td>5</td>
<td>File through filesystem</td>
<td>Files encountered as discrete set of items with associate paths and file names</td>
</tr>
<tr>
<td>4</td>
<td>File as “raw” bitstream</td>
<td>Bitstream encountered as a continuous series of binary values</td>
</tr>
<tr>
<td>3</td>
<td>Sub-file data structure</td>
<td>Discrete “chunk” of data that is part of a larger file</td>
</tr>
<tr>
<td>2</td>
<td>Bitstream through I/O equipment</td>
<td>Series of 1s and 0s as accessed from the storage media using input/output hardware and software (e.g. controllers, drivers, ports, connectors)</td>
</tr>
<tr>
<td>1</td>
<td>Raw signal stream through I/O equipment</td>
<td>Stream of magnetic flux transitions or other analog electronic output read from the drive without yet interpreting the signal stream as a set of discrete values (i.e. not treated as a digital bitstream that can be directly read by the host computer)</td>
</tr>
<tr>
<td>0</td>
<td>Bitstream on physical medium</td>
<td>Physical properties of the storage medium that are interpreted as bitstreams at Level 1</td>
</tr>
</tbody>
</table>

These levels go beyond Knowledge Organization Systems (KOS): “used to organize documents, document representations and concepts” (Hjørland 2008, 86). They act as preparation to care for the integrity in the storage systems. This is a very important first step to maintain the integrity of data, records, documents, and information until it is transformed into knowledge. In this regard, Digital Forensics exists as the first step to identify digital objects before later being integrated with Archival Diplomatics as a means to achieve knowledge. This knowledge has two meaningful uses: in certain legal or juridical cases and/or for archival or stewardships processes.

5.0 Concluding remarks

Digital Forensics research is currently being analyzed and explored, with interesting outcomes in several centers of research, law enforcement agencies, and universities. We hope that the knowledge brought by this science continues to increase for the benefit of digital records conservation. The concepts, practices, and tools of Digital Forensics are being applied in several knowledge fields, and we can notice the benefits this offers to disciplines, such as Information Science, Archival-Diplomatics, and Knowledge Organization.

A series of software products are being developed to carry out this work. One such example is “BitCurator,” a software produced in 2016 in a partnership between the

---

2 https://github.com/bitcurator/bitcurator-nlp/wiki
School of Information and Library Science at the University of North Carolina at Chapel Hill and the Maryland Institute for Technology in the Humanities. This software runs with natural language processing (NLP), developing for collecting institutions to extract, analyze, and produce reports on features of interest in text extracted from born-digital materials contained in collections (BitCurator 2018).

The knowledge of archivists is key to the arrangement and description of paper records. Now in the digital device era, in addition to the Archival-Diplomatics Knowledge, tools and the application of the techniques produced and developed by Digital Forensics are fundamental because they deal directly with the integrity, reliability, and authenticity of digital records. Identifying, seizing, imaging, and analyzing material, such as floppy disks, cassette tapes, compact disks, USB, and hard drives, could be better performed through the application of technological tools and digital forensics processes to maintain the data integrity.

References


Committed to a Narrative
Expressions of Knowledge Organization at The Henry Ford
Museum of American Innovation

Abstract:
Innovation is one of the most pervasive rhetorical tropes in information industries and technologies. How can we deconstruct the meaning and social currency of innovation? Although science and technology studies critics have approached this question in recent years, it deserves closer study from knowledge organization (KO) domains. The history of innovation is a history of classification. In the Library of Congress, the subject heading “innovation” is classed by industry: technological innovations, agricultural innovations, and organizational innovations, for example. This classificatory structure fulfills capitalist expectations of industrial economics and mystifies the material production of industries and technologies. Significantly, innovation also allows documentary institutions to periodize their materials and order objects of knowledge along a hagiographic trajectory. An important site of study for this issue is The Henry Ford Museum of American Innovation in Dearborn, Michigan. This project explores how all The Henry Ford’s diverse collections are classed as “innovative.” My research question asks how a priori commitments, particularly a narrative of innovation that intertwines technological and social progress, shape subsequent expressions of KO as found in the description and arrangement of museum artifacts? This question will guide my analysis of a museum that places the Rosa Parks bus, Civil Rights Movement memorabilia, and slave shackles in the same physical space as Thomas Edison’s electric pen, a Macintosh 512K personal computer, and early television sets. All these objects are classed along a singular trajectory—they are all hagiographic exemplars of American history. Central here is the argument that The Henry Ford classifies innovation by privileging its collection objects’ modes of inscribing meaning in a “cluster of mutually defining” (Gitelman 2000) technological and social practices of change that are themselves already privileged in canons of American history. In this way, we see how KO is not an idealized neutral space to judge and describe objects, but rather expresses a positionality from which regimes of KO are constructed.

1.0 Introduction
Innovation is one of the most pervasive rhetorical tropes in information industries and technologies. How can we deconstruct the meaning and social currency of innovation? Although science and technology studies critics have approached this question in recent years, it deserves closer study from knowledge organization (KO) domains. In the Library of Congress, the subject heading “innovation” is classed by industry: technological innovations, agricultural innovations, and organizational innovations, for example. This classificatory structure fulfills capitalist expectations of industrial economics and mystifies the material production of industries and technologies. Henry Ford is arguably the most canonically important industrial capitalist in American history. He is also celebrated as a legendary innovator, not only for his vision of industrial technologies (the assembly line) but also his bold social vision. His legacy as an innovator is enshrined at The Henry Ford Museum of American Innovation and Greenfield Village in Dearborn, Michigan1. The museum and historical village date to 1926, when he purchased 260 acres near his sprawling River Rouge Factory complex with the purpose of

---

building an equally sprawling museum complex filled with historical buildings and artifacts of domestic and industrial life (Swigger 2014, 17). Today, the space collectively known as The Henry Ford is the largest indoor-outdoor museum complex in the United States. This paper addresses the meaning-making of collections at the museum specifically, whose collection mission is to “illustrate the process and context of innovation.” Of critical importance, then, is an examination of how all the museum’s diverse artifacts and documents are classed as innovative.

This project examines how the groupings of collections at The Henry Ford express contingent meanings and relationships, and how these relationships constitute a particular ideological narrative of innovation. This paper demonstrates the fundamental positionality of any knowledge organization system. A popular replica and exhibit space at The Henry Ford’s Greenfield Village—Thomas Edison’s Menlo Park—allows a glimpse into how the museum positions innovation as a verb and noun in order to make a coherent narrative. This narrative, in turn, relies on tropes of American resilience and novelty. What follows is a discussion of other information and communication technologies—specifically a collection of those meant to evoke feelings of personal nostalgia—on display in the museum. The final section analyzes how The Henry Ford has evolved to communicate more directly with its local constituents in Dearborn and Detroit through the acquisition and display of American Civil Rights movement artifacts. Edison’s laboratory and twentieth century information technologies may not seem to share a common meaning of “innovation” with the museum’s Civil Rights exhibits and the infamous Rosa Parks bus, but as we shall see, they work in concert to collapse the meaning of innovation in order to serve a whiggish American history trajectory. Ultimately, The Henry Ford must maintain narrative control in order to represent and organize American history—a serious knowledge organization task.

2.0 The method

A central goal of this project is to bring together more closely media history and KO literature. If, as Lisa Gitelman and others have suggested, we can think of media as material objects and complicated structures of sociocultural communication (Gitelman 2008, 6), we should examine more closely how meaning-making takes place at the intersections of physical presence and ideas. Lorraine Daston has further emphasized the indeterminate slippage between the “stolidly functional things…[and how they] radiate an aura of the symbolic” (Daston 2007, 19). Such an acknowledgment is particularly important for understanding the conceptual and physical authority of KO systems.

This paper acknowledges the prodigious body of literature on Ford’s Greenfield Village, a strange and impressive village of preserved historic structures and built on site. Jessie Swigger’s foundational text “History is Bunk”: Assembling the Past at Henry Ford’s Greenfield Village (2014) built on archival evidence and secondary literature (such as biographies by Steven Watts and David L. Lewis) to explain how Ford built a selective genealogy of industrial and domestic habits and materials. The particular assemblage of actual historical buildings (including the transplanted Wright Brothers’ cycle shop and an actual Cotswold cottage transported from England) and altered spaces created by The Henry Ford’s staff over the second half of the twentieth century are a seductively idiosyncratic site of study, rich with anachronisms and impressive efforts at historical authenticity. The museum, which boasts 26 million artifacts spanning 300
years of history, is well-recognized across literatures of technology history, material culture, and public history. The Society for the History of Technology’s journal publication, Technology and Culture, has regularly featured articles by The Henry Ford staff and reviews of its exhibits and expansions. This project introduces this site of public and technology history to the domain of KO. Ultimately it argues that the act of “doing” public history is an act of knowledge organization.

In recent years the domain of knowledge organization has recognized museums as powerful sites for critical KO theorization and research. In her illuminating conceptual review on this topic, Hannah Turner emphasized how critical KO analysis of museums “posits museums as key sites of knowledge production and circulation and sets the stage for an understanding of the “background” work of museums as an important site for understanding knowledge organization more broadly.” (Turner 2017, 473). Turner and numerous other KO researchers including Rick Szostak (2017), Melissa Gill (2017), and Lala Hajibayova (2017) have all enriched the KO community’s understanding of how institutional infrastructures and sociotechnical affordances position regimes of knowledge and organization. I offer a contribution to this literature, and also expand its critical and methodological dimensions. By comparing the different ways The Henry Ford describes and models the world, this project asks KO researchers to confront museums collections’ ideological and narrative powers. It is specifically these powers that legitimize museums’ production and organization of knowledge.

Central to this project’s contribution to KO is the emphasis that museum artifacts act as documents with powers of inscription. This argument owes much to the documentality work done by Michael Buckland and Ron Day. Buckland’s foundational 1997 essay explained how museum objects’ semiotic and evidentiary mechanisms of meaning control their existence as documents (Buckland 1997). Day began his most recent book with an important question about Suzanne Briet’s formative 1951 conception of a document as an antelope as documented by scientists (as a stuffed type specimen, for example): “What is forgotten about particular beings when they are subject to (or subjects of) the representation of being, understood as essential universal types (i.e., as class members)?” (Day 2019, 3). What is lost when all the numerous artifacts on display at The Henry Ford are understood as exemplars of innovation? What does “innovation” not say about the RCA-Victor Console Television Receiver, or the Rosa Parks transit bus? These questions have important ideological implications for the organization of knowledge, where classes of documents act “not just descriptively, but prescriptively” (Day 2019, 1).

The intersecting domains of knowledge organization and museums must reckon with the full range of museums’ ideological powers. Kevin Coffee, Chief of Interpretation and Education at Lowell National Historic Park, outlined this extent by remarking that, “museums and similar cultural organizations have a fundamental function to define and control visual expressions of major social narratives.” (Coffee 2006, 435) The resulting “concentrations of ideological symbols” legitimize and reinforce certain narratives about society (Coffee 2006, 435). The organization of museum objects produce prescriptive knowledge about culture, society, and time. Moreover, we can only comprehend the full extent of this ideological operation by unveiling the moral orders that dictated the museum’s collections from its start (Woodson-Boulton 2007, 48-49). By making explicit the complicated lineage of The Henry Ford’s technological collections, we can better
understand why and how knowledge organizing commitments change in museum institutions.

3.0 Innovation and History

Physical exhibition spaces in museums act as a primary principle of division among artifacts. Greenfield Village contains the most popular exhibition spaces at The Henry Ford (visitors may purchase tickets for the museum, the village, or both). Menlo Park—an assemblage of replica structures and authentic artifacts of Thomas Edison’s experimental laboratory—was one of Ford’s first endeavors in creating Greenfield Village. Ford and his treasured Greenfield Village architect, Edward J. Cutler, recreated the building in 1929 with some original elements (https://www.thehenryford.org/collections-and-research/digital-collections/artifact/179489). As a canonical American inventor, we can look to Thomas Edison to better understand how the institution of The Henry Ford defines innovation. The assemblage of replicas and authentic artifacts cooperate to materialize a convincing general class of innovative objects.

The curious mixture of real and imitation at Menlo Park evince Henry Ford’s presentist image of historicity. Ford’s collecting commitments, which began with the relocation and restoration of historic structures at his ambitious Greenfield Village, hinged on his ambivalent nostalgic attitude, which was a reaction to the mixing of races and cultures forms in American cities. It is cruelly ironic that his pre-urban, pre-industrial nostalgia was in large part due to the changing economic conditions of industrial cities that Fordist capitalism engendered. Ford’s ideology of innovation is best summarized in his infamous 1916 remark that “History is more or less bunk. It’s tradition. We don’t want tradition. We want to live in the present, and the only history that is worth a tinker’s damn is the history that we make today.” Ford’s establishment of Greenfield Village and the then-termed Edison Institute and Industrial Museum (Swigger 2014, 3) operated with these two connected goals: to celebrate a hagiographic American history of industrial progress, and to root acts of progress “today” in these whiggish lessons of the past. This narrative control is ideology—specifically, an ideology of American innovation.

The ideological power of Greenfield Village and Menlo Park lies, in part, in how convincing the environment is. The combination of replica and original produces an effective model of historic reality. When Edison himself toured the site with Ford in 1929, he remarked that the site was “99.9% perfect” (Swigger 2014, 78). The Henry Ford even classifies the space itself as an artifact entity, with typical catalog attributes such as creation date (1929, the year Ford reconstructed the building), subject date (1876-1883, the years Edison actively used the laboratory), and materials (wood, glass, metal). In addition to environment-setting objects (furniture, laboratory equipment), the interior contains many of Edison’s original patent models, including his successive printing and duplex telegraphs, electric lights, phonographs, vote recorders, and the electric pen. The artifacts are not organized in a linear progression, but instead are arranged in-situ, placed in naturalistic settings around the interior. Despite the naturalistic physical framing, however, they are described (in both written collection descriptions and scripts enacted by historical reenactors) as a succession of increasingly efficacious and influential inventions. In reality, Edison’s electric pen was a commercial failure and evidence of the nuanced social negotiations of Edison’s many inventions (Gitelman
2000, 5); however, The Henry Ford does not represent the artifact in this way. It is instead classified as an innovation in the overarching trajectory of Edison’s innovation.

In this way, The Henry Ford defines innovation as a verb and a noun. As a verb, the institution follows the normative conception of what Benoît Godin and others have defined as a “linear model of innovation” (Godin 2006; 2012; 2015). While there are numerous ways of thinking about models, it is most applicable here to conceive of models as narratives. The linear narrative of innovation consists of three discrete stages: a research phase, an applied research and development phase, and a final phase of production and dissemination. This narrative is a simplified representation of segments of reality, or potential realities, and does not (or cannot) account for the complexities of scientific practice, market forces, and supply and demand. In this respect, the narrative of innovation is a concentrated set of symbols—it is ideological. Furthermore, this suggests how we can understand the narrative properties of classification and knowledge organization.

4.0 Communication and Media Technologies, Communicating and Mediating Innovation

It should not be surprising that the tens of thousands of objects exhibited at The Henry Ford communicate meanings relationally. As a whole, the museum objects are classed as innovative. Edison’s electric pen is defined as innovative in relation to the other objects in the reconstructed Menlo Park; the assemblage of objects and structures that comprise the Menlo Park site evoke existing grand narratives about Edison as the grand American inventor. As the largest indoor-outdoor museum complex in the United States, it is helpful to examine how objects in the museum proper (as opposed to Greenfield Village) communicate meanings of innovation. Doing so demonstrates how The Henry Ford manifests an organized physical landscape of knowledge comprised of mutually defining media.

The permanent exhibit “Your Place In Time” situates the viewer differently than Menlo Park and Greenfield Village. The exhibit space is structured as a chronological journey through twentieth century popular American technologies. It contains objects of immediate personal nostalgia for viewers, from cassette and record players to an interactive MTV music video creation station. Neither the exhibit space nor its artifacts immediately communicate a message of American innovation the way Menlo Park does. Significantly, however, many of the Your Place In Time artifacts are included in the same collections class as Thomas Edison’s inventions. The Macintosh 512K Personal Computer, for example, is classed alongside the Edison electric pen in the museum’s “Information Technology & Communications” digital collection set. In this example, The Henry Ford partially flattens the historical differences between these artifacts to allow their mutual inclusion in an object type category. Here, the definition of “innovation” as a noun takes precedence over the nuanced negotiation of technology history. After all, visitors at the physical “Your Place In Time” exhibit do not bear witness to a reenactment of the innovation process, the way they do in Menlo Park and Greenfield Village. They instead see and interpret inventions—the end-products of the linear model of innovation.
The different viewer positions at Menlo Park and inside the museum—where the former invites viewers to witness reenactments and engage with physical spaces and artifacts, and the latter invites viewers to look at artifacts and identify with them based on their own personal experiences—create slippages of meaning for understanding innovation. As the viewers visit the different exhibit spaces at The Henry Ford, they engage with narratives that classify innovation as a thing and a process. In the next section, we will see how The Henry Ford additionally classifies innovation as a trait held by a person.

5.0 Rosa Parks, Innovator

The Henry Ford’s evolution should be contextualized in a larger trajectory of public history and technology museums. As technology historians have explained, the field began to question its pervasive and deterministic narratives as early as the 1980s (Staudenmaier 2002, 168-181). The Society for the History of Technology concretized more critical and reflexive methodologies during a 2007 workshop sponsored by the organization and the National Science Foundation. Colin Divall and David Edgerton both began publications at this workshop that asked the field to mobilize a more critical, conceptual framework for the domain (Divall 2010; Edgerton 2010). Edgerton in particular critiqued existing scholarship for focusing too myopically on technological novelties—innovations. Edgerton urged scholars to analyze materials that communicate something about popular understandings of technologies as they emerge in socially, politically, and geographically situated times (Edgerton 2010). Applying this more dynamic imperative to public history, Divall quoted past president of The Henry Ford Harold Skramstad: “...the fundamental challenge is to design exhibitions that have a clear and coherent intellectual intent while at the same time providing engaging individual experiences” (Skramstad cited in Divall 2010, 957). In other words, technology history museum professionals must recognize their own agency in creating complex analyses of technologies, which then inform audience experiences. This co-productive turn toward museum and audience dialogue has affected other public history sites in the United States such as Lowell National Historic Park (Goldstein 2000, 129-137). It is no mere coincidence that The Henry Ford publicly expanded its civil rights artifact collections beginning around 2001, the same era these conversations engaged technology historians. It is also no mere coincidence that this focus on civil rights artifacts coincided with increased support for Greenfield Village’s African American Family Life and Culture program and partnership with the emerging National Arab American Museum (in 2000 Dearborn’s population was more than 28% Arab American; Detroit’s population was more than 80% Black).

As The Henry Ford changed their commitments to meet evolving community needs, their collections came to include an American Democracy and Civil Rights focus area. In 2001 The Henry Ford purchased and restored the “Rosa Parks Bus”—the Montgomery, Alabama city bus on which Parks initiated the Montgomery Bus Boycott in 1955. The bus is the apex holding of The Henry Ford’s permanent “With Liberty and Justice for All” exhibit. An oft-reproduced Pete Souza photograph of Barack Obama sitting alone in the bus in 2012 demonstrates the material and visual power of the artifact (Peralta 2012). Parks holds a distinguished position in the museum’s webpage list of “historic innovators” (as opposed to the museum’s contemporary list of innovators, which
includes figures like Bill Gates, Elon Musk, and Steve Wozniak)—besides Parks the list features Henry Ford, George Washington Carver, Thomas Edison, and the Wright Brothers.

The museum also now classes innovation by certain social behaviors or “habits”: collaboration, breaking rules, learning from failure, remixing, and being curious (https://www.thehenryford.org/explore/stories-of-innovation/visionaries/). All of these classes deploy a certain linear rhetoric of social progress—they infer new relationships (collaboration, remixing), a progression of standards (breaking rules, learning from failure), and new applied thought processes (being curious). This classification of innovation is socially prescriptive, but ethically and morally ambivalent. The Henry Ford describes Parks as a specific kind of innovator, and one that distinguishes her immediately from the likes of Edison and the Wright Brothers: “…her simple, spontaneous act embodies the notion of social innovation—that a new idea or way of doing things can have such far-reaching impact, that it renders old ways obsolete and radically alters how people think about themselves, their social interactions, and their place in the larger world.” (https://www.thehenryford.org/explore/stories-of-innovation/visionaries/rosa-parks/)

How is this description similar or dissimilar to the museum’s classification of innovation by behavior type? The most appropriate class to categorize this description of Parks is “breaking rules.” Her act of protest “radically” changed others’ behaviors in a new fashion. What about collaboration? This descriptive web page text notes Parks’ attendance at the Highlander Folk School civil rights training during the summer of 1955, but on the whole it canonizes Parks as a unique visionary of the Civil Rights movement: “Many consider her singular act of protest to be the event that sparked the Civil Rights movement […] her flawless character, her quiet strength, and her moral fortitude caused her act to successfully ignite action in others.” (https://www.thehenryford.org/explore/stories-of-innovation/visionaries/rosa-parks/) The text’s uncited quotation of the iconic phrase “they had messed with the wrong one” additionally suggests a move away from failed protests or tactics heretofore in the Civil Rights movement—an echo of The Henry Ford’s innovative behavior class “learning from failure.” Curiosity and remixing are similarly inferred, with the former attributed to her marriage to Raymond Parks and subsequent exposure to the Civil Rights movement, and the latter attributed to her combination of extraordinary moral characteristics. This example shows how The Henry Ford uses a flexible classification structure to appropriately situate the creators of its disparate collections.

6.0 Conclusion

The Henry Ford Museum of American Innovation has evolved significantly since Henry Ford’s death, both in terms of its mission and collecting principles. The broad class of “innovation” has allowed the museum to expand its collections and exhibits to meet the evolving needs of its surrounding communities and the evolving expectations of technology history and museum studies. By comparing how the museum accounts for disparate definitions of “innovation,” we can see just how flexible the museum’s ideological classification needs to be. At the same time, this classification flattens the ideological differences between these figures and unites them on a singular trajectory of American history. This paradoxical classing—where innovation means something very
temporally coherent, and so many characteristics that it means little at all—allows us to question the ideological power prescribed by museums.

Meaning in museum spaces takes place at the negotiation between the institution and viewer agency. The settings for these interactions—whether they are assemblage spaces of authenticity and replica, or more traditional museum exhibit spaces—play a significant role in what meanings are produced. At The Henry Ford, institutional agents must account for highly disparate settings (a historic village and interior museum) and collections of artifacts. Under the broad, flexible, and at times contradictory class of “innovation,” The Henry Ford must bring together these settings and artifacts. To do this effectively, the institution projects a grand narrative of American history. Narrativity is key here. In the organization of historical moments, narrativity creates a necessary flow. As Ron Day recently explained, “The theoretical construct of the past as continuous, much less returnable, is an explanation that depends on narrative, historiographical, conventions…Time must be seen as continuous in order for component parts to be retrieved from its series” (Day 2019, 106). Sites of public history and technology, such as The Henry Ford Musem of American Innovation, offer complex opportunities to unveil these narrative—and ultimately ideological—mechanisms. As KO research continues to examine such sites, we should critically consider the fundamental positionality of any KO system.

References


Terminological Relations of a Thesaurus for University Cultural Infrastructure Terms

Abstract:
The objective of this work is to define the terminological relations of the University’s cultural infrastructure with the purpose of developing a thesaurus, in Spanish, that contributes to the thematic organization of a database. This will be used for a better communication and understanding among organizers of cultural activities in a university environment. In order to achieve this project, we used a descriptive method integrating techniques of observation. We completed interviews and analyzed the existing literature by using the factual records as instruments. Different worksheets and a questionnaire were also used to obtain a domain consisting essentially of cultural spaces such as auditoriums, libraries, cinemas, esplanades, outdoor forums, museums, concert halls, conference rooms, audiovisual projection rooms, multipurpose rooms, dance halls, music halls and theaters, in addition to their relationship with technical resources and areas and attributes. Thus, obtaining an arborescent structure for the development of a thesaurus destined for the university environment.

1.0 Introduction

The purpose of this study is to define the terminological relationships on a university cultural infrastructure, in order to develop a thesaurus, in Spanish, that contributes to the thematic organization of an information system. The thesaurus will be used to achieve a better understanding and communication among the organizers of different cultural activities in a university environment of large dimensions——one that has many campuses in a same country or even in others (Universidad Nacional Autónoma de México. UNAM)

The organization of cultural activities in the university implies an acquaintance with the spaces, the location and the technological resources that each enclosure possesses. In this way, deciding the correct place to carry out the diverse activities becomes easier. However, this paper focuses on solving this communicative gap through organized and representative terms. The comprehension of the field of action and the elements that characterize it implies an intellectual effort that requires an in-depth study of conceptualization though terms that represent and directly interrelates them.

University cultural spaces are of two types: professional spaces for a given activity or multipurpose spaces. The acts to perform in the venues are, but not limited to, presentation of plays, concerts, dance shows, screening of films or videos, university protocol events such as awards, conferences, seminars, with each of these actions including different supporting elements.

1 This is part of a project supported by the PAPIIT IT400318 resources
Thus, in the communication between the various university users of cultural spaces it is essential to use the terminological designations that are understandable, both for organizers, and the participants in the diverse acts in order to avoid the errors that human communication can cause. “The purpose of a thesaurus is to guide the indexer and the search engine to select the same preferred terms or combine the preferred terms to represent an assigned theme” (ISO 2011, 12).

The concept of indexing language to represent documents in databases is used in Information Science, not only to index documents but also to structure the database (Fugmann 1992).

2.0 Methodology
To develop the thesaurus a combination of different methods was used: description (scouting) of the spaces and their infrastructure, interviews with decision makers and technicians, design and testing of the questionnaire, analysis of terminology in interviews, content analysis in literature on cultural venues, comparison of the terms in specialized thesauri in the topics. All this was done to gain knowledge of the domain in which a thesaurus operates; “thesaurus was on the agenda, but the design was to be based on the results of the domain study” (Likke 2001, 774).

2.1 Art & Architecture Thesaurus (Getty Research Institute)
The consultation of thesauri related to the subject before undertaking the construction of a thesaurus is mandatory. In the case study on cultural venues, the most representative is the Art & Architecture Thesaurus (AAT) of the Getty Research Institute. This thesaurus has a published translation in Spanish; however, we observed that the approach is centered in the Anglo-Saxon context with hierarchical and faceted terms related to art, architecture and other material cultures, associated concepts, periods and activities, etc. Both the hierarchies used and the terminological approach are not focused on a university environment, where the thesaurus operates from, since the AAT reflects terms linked to the arts, the entertainment and shows corresponding to a wider cultural spectrum.

The relations in AAT are wide and confusing for the communicative situation that we are trying to solve with the thesaurus that we are building. This thesaurus can then be a contribution to transliterate Anglo-Saxon terms in an Ibero-American context.

“The need for the definition of an international “glossary” for architectural heritage and by extension, for cultural heritage has arose. This need has originated from the fact that there are various methodologies regarding heritage documentation. Various vocabularies and thesauri are used in the field of conservation, while the variety of “uniqueness” of each cultural artefact turns its categorization into a difficult endeavor. In addition, not only spatial information needs to be standardized, but also the related metadata. Multilingualism, the translation of terms and the existence of many local words for the description of the same object, are the most important challenges when structuring vocabularies. Therefore, the attempt to describe an object with terms understandable to every culture and the adoption of a common “linguistic ground” meets a number of difficulties” (Maietti 2018, 107).

The purpose of the thesaurus to be developed, unlike Paul Getty's thesaurus, is to group and organize the terms that describe cultural spaces, technical resources, areas and attributes as the main categories from which other ramifications are derived. These necessities make unbeatable differences not only in the main categorization, but in the terms that represent the concepts in a university cultural environment that does not include mass entertainment, for example, at a cultural level.
In this way, the differences and needs in the terminological structure must be remedied from different disciplinary contributions since a single approach is not enough to explain social, natural, economic and cultural phenomena. Hence, it is necessary to broaden the horizon with conceptual elements representative of the interests in indexing and information retrieval (Hjørland 2002a). “It is really important to know the most important information sources in one or more domain at a rather detailed level. It has a strong relevance for practical information work” (Hjørland 2002b, 425).

2.2 Definition of the University Cultural Infrastructure

Because the purpose of this thesaurus is to organize the terminology of an information system, the next step was to define and understand the concept that would properly name the information system that was developed. After studying and analyzing the scope of operation, University Cultural Infrastructure was chosen.

Infrastructure can be understood, in a general aspect, as the set of properties and resources that some individual, company, or institution has. Nonetheless, there are certain definitions from institutions and initiatives involved in national cultural development, in the Anglo-Saxon world.

For the government authorities in London, cultural infrastructure is the grouping of creative work sites, performing arts rehearsal spaces, music recording studios, film and television studios (Khan 2019). While the national organization dedicated to cultural development in Canada, considers that cultural infrastructure is comprised of resources and spaces built specifically or adapted for use. Examples of the spaces that are part of the cultural infrastructure are the performing arts centers, galleries and museums. (CCNC Special Editions 2009).

For the Australian government the cultural infrastructure is considered as the buildings built or acquired to create, share and enjoy the artistic and cultural activities, such as theaters, galleries, museums, libraries, archives, community rooms, cinemas, public art and spaces for outdoor events. (Create New South Wales 2020)

From the information presented above, it can be deduced that, in Anglo-Saxon countries, cultural infrastructure refers specifically to movable and immovable property that are conserved, acquired or adapted for performing arts such as dramatic performances, music and dance, such activities are carried out —— usually in dance halls, concert halls, auditoriums, theaters, galleries and museums. In a university environment the cultural infrastructure also includes other types of spaces that are used to carry out activities related to the academic life.

For the purposes of this work, the University Cultural Infrastructure (ICU) is defined as all those cultural spaces where artistic and cultural activities such as dance, theatrical performances, film projections, conferences, concerts, art shows, etc. are held and where technical resources —— movable goods; the set of tools, instruments and artifacts used to perform cultural activities—— are required. These resources can be specialized and used for various events.

2.3 Exploration of spaces and resources

A first approach to the domain that would comprise the database and the terminology used was defined by observing the cultural enclosures and their components, supported by the resulting interviews. We classified the types of spaces based on the resources that
characterize them and we used an initial terminology compared to the definitions in specialized works such as glossaries, dictionaries, encyclopedias, controlled vocabularies, university academic works, videos, and images. The plans of the university units were also consulted to obtain different data. Thus, we selected and discarded the consulted works and the digital contents based on the level of description of the distinct entities and also on their compatibility with the ICU.

We developed a questionnaire to interview the authorities and managers of university spaces, that was first tested in certain representative venues. Additionally, we compared it to other questionnaires applied in cultural information systems of other national and foreign institutions. Based on this, we visited the university units in Mexico City and in Mexico’s different states where diverse facilities are owned. This questionnaire was also sent to dependencies abroad in order to also include them in the information system.

As a result of the combination of the descriptive and exploratory methods and their application to the definitions of the existing enclosures, we carried out a terminological integration to form a specialized vocabulary. This allowed for a first categorization, typifying the cultural spaces to develop the domain trees in separate groups: auditorium, library, cinema, esplanade, outdoor forum, museum, concert hall, conference room, audiovisual projection room, multipurpose room, dance hall, music room, theater.

2.4 The ICU domain trees

After having reviewed and selected the ICU domain terminology, domain trees were designed for each cultural enclosure. This allowed us to identifying the differences and similarities between the elements that compose the different spaces. The first venue analyzed was the theater, essentially a cultural space par excellence, therefore, it was the starting point to determine the terminology of the rest of the cultural spaces that were similar, such as the auditorium, the cinema, the concert halls, the halls of conferences, etc. These different enclosures have components and elements in common (room, armchairs, hallways, etc.) the differences lie in the architectural design and the alterations in scale, in addition to the activities that are carried out in these cultural spaces.

Although the university’s cultural spaces include enclosures and multipurpose sites to develop the activities, it is necessary to distinguish each of them both for their particular characteristics and their technical resources. This is achieved through visiting and being acquainted with the distinct enclosures, observing their characteristics and properties based on the terminology that represents each one of them. Some terms are usually specific to an area; in that case these terms become the qualities that clearly distinguish some areas from the rest.

An arborescent structure should avoid, as much as possible, the repetition of words and denominations. Consequently, two cultural spaces should not be in the same field or level since each one is different for its attributes and for the activities that are carried out in them.

Domain trees must be based on facts and documents that prove the existence of the terminology; however, there is subjectivity bias in the interpretation of the contents and the lexicons. “The pragmatic approach to classification through meaningful units of knowledge must be based on recognition of the obvious truth that any single unit may be meaningful in any number of different relationships depending on the immediate purpose. Thus, it is the external relations, the environment, of the concept that are all-
important in the act of classifying… Relationship is not a universal, but a specific fact unique to the things related, and just as these relations reveal the nature of relata, so the relata determine the character of the relationship. (Shera 1951, 83-84)

Shera’s statement is perhaps the argument that justifies the need to elaborate a special thesaurus to demonstrate the reason why a pragmatic classification highlights the necessary properties of a concept, in a specific information system. The relations that are developed for an information system are disimilar from those that stand out for another system whose objectives are different. (Kwaśnik 2019)

<table>
<thead>
<tr>
<th>AUDITORIUM</th>
<th>THEATRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HALL</td>
<td>HALL</td>
</tr>
<tr>
<td>AISLE</td>
<td>AISLE</td>
</tr>
<tr>
<td>HALL STAIRCASE</td>
<td>HALL STAIRCASE</td>
</tr>
<tr>
<td>HALL RAMP</td>
<td>HALL RAMP</td>
</tr>
<tr>
<td>SEATS</td>
<td>SEATS</td>
</tr>
<tr>
<td>NUMBERED SEATS</td>
<td>NUMBERED SEATS</td>
</tr>
<tr>
<td>ROW</td>
<td>ROW</td>
</tr>
<tr>
<td>FIRST FLOOR</td>
<td>FIRST FLOOR</td>
</tr>
<tr>
<td>SECOND FLOOR</td>
<td>SECOND FLOOR</td>
</tr>
<tr>
<td>BOXES</td>
<td>BOXES</td>
</tr>
<tr>
<td>MEZZANINE</td>
<td>MEZZANINE</td>
</tr>
<tr>
<td>ORCHESTRA</td>
<td>ORCHESTRA</td>
</tr>
<tr>
<td>PLACES FOR DISABLED</td>
<td>PLACES FOR DISABLED</td>
</tr>
<tr>
<td>STAGE ACCESS</td>
<td>STAGE ACCESS</td>
</tr>
<tr>
<td>LOBBY</td>
<td>LOBBY</td>
</tr>
<tr>
<td>VERTICAL CIRCULATION</td>
<td>VERTICAL CIRCULATION</td>
</tr>
<tr>
<td>ELEVATOR</td>
<td>ELEVATOR</td>
</tr>
<tr>
<td>STAIRCASE</td>
<td>STAIRCASE</td>
</tr>
<tr>
<td>EMERGENCY STAIRCASE</td>
<td>EMERGENCY STAIRCASE</td>
</tr>
<tr>
<td>PUBLIC TOILET</td>
<td>PUBLIC TOILET</td>
</tr>
<tr>
<td>BAR</td>
<td>BAR</td>
</tr>
<tr>
<td>CAFÉ</td>
<td>CAFÉ</td>
</tr>
<tr>
<td>RESTAURANT</td>
<td>RESTAURANT</td>
</tr>
</tbody>
</table>

Figure 1 Comparison between the domain trees of the audience spaces in a theater and an auditorium. Most of the similarities between the terms that name the spaces were found in their public areas. Thus, it became evident that we could not simply classify the spaces based on their type of enclosures, rather we had to begin to classify them based on the different relationships with the areas into which they are divided.2

On the one hand, in the case of outdoor spaces, such as open-air forums and esplanades, we identified that these have only very few elements that support the realization of the cultural activities. In this way, despite being open places they must have a specific category that identifies them. Hence, it is important to categorize their elements in order to link them with the technical resources that support outdoor presentations, such as concerts, fairs, exhibitions, festivals, etc.

On the other hand, there are spaces that are part of larger architectural ensembles. For example, museums, in their architectural program, can include, in addition to exhibition spaces, a library and even an auditorium. In this sense, to repeat the arborescent structure of a museum’s auditorium or library would be redundant.

---

2 There are certain terms used in Spanish that do not have a clear equivalence in English; there are not definitions or names for specific parts of auditoriums and theaters.
The close relations that keep within them a set of the different cultural spaces, more importantly, keep the similarities between the elements that compose them and the resources that are integrated in them. These diverse elements showed that there was a necessity to organize the general arborescent system in a different way. That is to say, the organization did not have to be based on the different existing cultural spaces of the University. We had to relate them in such a way that the characteristics and resources, of each of the spaces, would be distinguished through the sum of what they do and do not possess in their exteriors and interiors.

The main categories should have the conceptual capacity to characterize the spaces by their similarities and differences without repeating the component elements of each space, because this would also be useful when programming, for example: an auditorium that has external communication on stage, such as theaters. By including this auditorium related to an area outside of the stage, it shows the ability to introduce elements into the auditorium that can make it function as a theater. It was necessary to create a simple categorization that highlighted the common elements in different spaces that worked to bring programming possibilities together and did not repeat the common characteristics in each space.

Additionally, having developed an arborescent structure for each cultural space allowed us to understand its composition and elements to establish the appropriate connections between the different categories, classes, and subclasses, in order to create a level of integration in the terminology. For this purpose, a conceptual universe was created. This includes three categories: A) Cultural spaces, B) Technical resources and C) Areas and attributes within the domain of the ICU. Some of these elements are closely connected depending on whether they are linked according to their degree of affinity.

![Figure 2. Conceptual universe of the thesaurus of the ICU and its three main categories.](image-url)
3.0 Discussion and results

Each category contains an arborescent structure (which includes related elements) that is associated with other thematic ramifications, but without losing its own hierarchical composition. In this sense, associations are the way in which the terms of each category are intertwined. In this sense, the related terms (RT) can cover the relationship “from many to many” elements to connect them to the categories and they also cover the relations “from one to many” to link the associations that are shared in more than one term. Specific terms, however, (ST) with a “one-to-one” principle are also necessary, according to the type of link between space, resources, areas, and attributes.

The category of cultural spaces (A) is integrated by all those places where cultural activities are programmed and carried out. Subject to this category are the classes and subclasses that describe the architectural elements that are understood as the fixed spaces physically delimited within a general area or space and within the components of the architectural elements; this can be fixed or semi-fixed and support the realization of the main activity, depending on the space in question. Below, we present a diagram of the connections or associations of the main three categories. We specifically highlight the Cultural Spaces because they are the ones that have both the areas and the places where technical resources are used.

![Figure 3. Relations between the general elements of each category](image-url)
3.1 Evaluation

Once the proposal was developed and inserted into the software, the related and hierarchical descriptors were obtained. The indexes created with the help of the software were handed to the Cultural Diffusion Coordination of the UNAM which is directly responsible for evaluation of the information system. After their evaluation, with the help of their observations, the thesaurus’ final validation will proceed. As for the synonymy relations and the meanings of the descriptors, they were incorporated in some cases necessarily to present the structure in a simpler way to the evaluators who would be in charge of the operation of the information system.

“ The relationship of intersection refers to the relationship that the meaning of one word intersects with the meaning of the other word to a certain extent. In this case, the two words are at the same level. There is no upper term, nor lower term, which is the case in the previous relationship. Accurately speaking, most semantic synonyms are also in this relationship. The difference between intersected semantic synonyms and intersected contextual synonyms lies in the fact that the intersected part of semantic synonyms is the whole part of one meaning of the synonymous words, whereas the intersected part of contextual synonyms is only a part of one meaning of the synonymous words” (Zeng 2007, 35).

Although the categories are grouped according to their common elements, through intersection, the relationships that are created to give functionality to the thesaurus can reflect a greater sense of complexity as is the case with the synonyms of the domain. Since in an abstract sense, they share the same level despite the order of preference by the community. Synonyms, then, were not included in this first stage because the activity organizers can express their preferences without feeling influenced by other contexts.

“ As mentioned above, a core problem in IR is the adequate “mental modeling” of subject literatures. What categories and concepts are we talking about? In interacting with subject literatures, users are interacting among other things with

1. Different kinds of knowledge fields with different social and cognitive organization.
2. Different languages for special purposes (LSP)
3. Different kinds of research methods
4. Different kinds of, among other things, primary, secondary and tertiary documents
5. Different patterns of cognitive authority.

Although in this paper we only present the hierarchical and associative relationships, in the thesaurus, which is already elaborated, there is a more complete version that includes synonymy relationships, the meaning of the terms, and the scope notes that define the meanings of the terms in the field of operation. Through the university body in charge of Cultural Diffusion, composed of specialists in the field responsible for the enclosures, technicians are being consulted to validate and obtain a quality intellectual product, capable of responding to the needs of the communities concerned.

4.0 Conclusions

The three main categories in which the terminological relationships of the University Cultural Infrastructure are organized were achieved after several conceptual approaches, particularly, after confronting and comparing the domain trees of the different facilities registered in the university environment.

The initial categories for organizing the thesaurus were the types of venues that the university noticed, after trying to relate them to each other, were not adequate. We understood that this was because the types of venues have similar characteristics and that
are not of help for their definition. The enclosures are better defined through the areas and the attributes that are related to them.

Domain trees are useful for starting to develop hierarchical relationships between terms; however, the thesaurus must present an integrating tree development that is not the sum of each domain tree.

Thesauri are specific structures in an area of operation because their conformation supposes facilitating, understanding, and communication by studying and reflecting the terms used by the community for which they are intended. They also reflect the characteristics that are of interest to highlight to fulfill the system's objectives.

The relationships around significant knowledge units are of a pragmatic classification; this is crucial to determine when structuring a thesaurus to give clarity to an information system.

The utility of the University Cultural Infrastructure Thesaurus is operative to organize the information system on university events, taking advantage of the spaces and resources that are available. Its conformation was a collaborative intellectual work to solve the communicative situation; we did this by conceptualizing and relating the characteristics of the environment where the information system operates.

References
Using the Concept of Warrant in Designing Metadata for Enterprise Search

Abstract:
Metadata is an issue of growing concern in enterprise search. Several authors argue that adding metadata can improve the findability of content (Cleverley and Burnett 2015; Schymik et al. 2015; Stocker et al. 2015; White 2016). This paper proposes to use the concept of warrant when designing metadata for enterprise search. The paper combines the concept of warrant (Barité 2018; Beghtol 1986) with other concepts to analyse eighteen articles in library and information science journals from the period 2000-2019. The articles report on the design of thesauri, taxonomies, classification schemes, metadata or ontologies in work domains. The results indicate that the main warrants used in the articles are information sources’ warrant, task performers’ warrant, and work context warrant. The warrants illuminate where and how the concepts and terms in the designed systems are grounded. We argue that the concept of warrant is useful to analyse and choose different perspectives when designing metadata, but that other concepts and frameworks are needed for evaluation and implementation of metadata in enterprise search.

1. The aim and scope of the study
This study is concerned with the design of metadata for enterprise search, and how the concept of warrant is helpful in this respect.

In many organisations today information is created and used during different work activities. Enterprise search is an area that occupies itself with the retrieval of this information. Enterprise search can be defined as “search of digital textual materials owned by an organization, including search of their external Web site, company intranet, and any other electronic text that they hold” (Hawking 2011, 641). Thus, enterprise search is connected to the retrieval of information in an organisation or a company, that is, the work domain (Lykke Nielsen, 2001). However, search of digital textual materials within a work domain is not straightforward. There is evidence indicating that employees spend much time searching for information needed to perform their work, often without finding it (Hawking 2011). Surveys indicate user dissatisfaction with enterprise search applications (White 2016). According to White (2016, 188) user needs for a better search performance “cannot be achieved by technology alone”. A potential way forward in this respect is the use of taxonomies and thesauri for adding metadata to content. Other authors also pointed at the need for metadata in enterprise search (Cleverley and Burnett 2015; Kruschwitz and Hull 2017; Schymik et al. 2015; Stocker et al. 2015).

As enterprise search is connected to the domain of work, it is relevant to look at design of metadata from a work domain point of view. The aim of this study is to analyse journal articles that report on the design of knowledge organisation systems (KOS) or metadata in the domain of work. Using the concept of warrant, the study categorises different approaches to the design, and discusses their relevance for enterprise search. The research question for the study is “What are the main warrants used for metadata design in the domain of work?”
2. Theoretical framework for the study

In the design of metadata or KOS for enterprise search the concept of warrant (Beghtol 1986) is useful. Warrants are originally “bases on which library classification schemes could be constructed” (Hider 2015, 156). According to Barité (2018, 518), Beghtol (1986) was the first to propose a general definition of ‘warrant’ that opened up for other contributions than the original literary warrant. Barité (2018, 528) provided an updated list of different warrants, and also described the categories literary warrant, user warrant, cultural warrant, academic warrant and organisational warrant. Summarising the two articles (Barité 2018; Beghtol 1986), literary and terminological warrant take the subjects in books or documents, and the terminology of a subject field respectively as the basis for classification schemes. Scientific warrant builds on scientific disciplines and agreement in scientific communities, and academic warrant builds on the opinions of experts. Educational warrant also builds on scientific disciplines. Institutional or organisational warrant has the needs of particular organisations in mind, and enquiry or user warrant the needs of particular users. Cultural warrant as a principle may lie behind different classification schemes.

Warrant is chosen in this study because it can illuminate the methodological level in the design processes, called for in previous research (Mai 2008). Warrant has recently been used by Hider (2015) to analyse and classify online schemas and vocabularies, based on their reported development methodology. Kwasnik (2010, 108) suggested that warrant can be used to understand, analyse, evaluate and design classifications.

Warrant has its origin in library and information science and the bibliographic domain. For the purpose of this study some warrants have been adjusted to fit the domain of work. Literary warrant was originally based on the subjects of books or documents (Barité 2018, Beghtol 1986). In the work domain ‘literary’ may not be sufficient to denote all relevant types of information on which a KOS can be based. ‘Information sources’ is a concept used in the conceptual framework for tasks (Byström and Hansen 2005) and in the cognitive work analysis framework (Fidel and Pejtersen 2004). This concept denotes information sources that are needed by task performers or actors. ‘Information sources’ warrant’ is in this study used for design that is based on (textual) information sources.

User warrant is based on the information needs of particular users (Lancaster 1986). In a work domain people can have more roles than ‘users’ of information, for example ‘actors’ (Fidel and Pejtersen 2004), that is someone that is involved in actions related to work. Another concept is ‘task performer’ (Byström and Hansen 2005), that is someone who performs a work related task. The information use environments model (Taylor 1991) focused on ‘sets of people’ such as for example the professions or the entrepreneurs. ‘Task performers’ warrant’ is used in this study for design based on people performing tasks or otherwise connected to a work domain.

Organisational warrant has its basis in the needs of particular organisations (Barité 2018, 539). As such it denotes a context around work tasks. But according to Taylor (1991, 226) “the organization is but one setting” that a group of people may work in. The cognitive work analysis framework (Fidel and Pejtersen 2004) has several dimensions related to work context: work environment, work domain, organisation, and activity. The conceptual framework for tasks (Byström and Hansen 2005) discussed...
contextual and situational attributes. For the purpose of this study ‘work context warrant’ is used, because it can include various aspects related to the context of work.

Summarised, for the purpose of this study, literary warrant is adjusted into information sources’ warrant; user warrant into task performers’ warrant; and organisational warrant into work context warrant. The warrants denote on what bases the various KOS have been designed.

3. Methods in the study

This study analyses journal articles that report on the design of KOS or metadata in the domain of work. The journal articles were purposively sampled, and analysed qualitatively (White and Marsh 2006). First, I carried out literature searches in the databases Web of Science and Library and Information Science Source. The search strategy is reported in Table 1. One category in the search strategy was design, and the other category was knowledge organisation system. The categories were represented by different search terms by using the OR operator, and combined by using the AND operator. I searched in the title fields of the two databases. I also added one known article to the results.

<table>
<thead>
<tr>
<th>Table 1: Search strategy for literature review</th>
</tr>
</thead>
<tbody>
<tr>
<td>(design* OR build* OR built* OR construct* OR develop* OR generat* OR creat*) AND</td>
</tr>
<tr>
<td>(thesaur* OR ontolog* OR vocabular* OR taxonom* OR classif* OR &quot;knowledge organisation system*&quot; OR KOS OR metadata*)</td>
</tr>
</tbody>
</table>

Next, I narrowed the search results down to articles in core library and information science journals (Tuomaala, Järvelin, and Vakkari 2014), in addition to Knowledge Organisation and Cataloging & Classification Quarterly, published in 2000-2019. I read the title and abstract of references, and in many cases also the full text of articles, in order to determine whether the article was relevant or not. The criteria for inclusion of articles are shown in Table 2. I included articles that reported on the design of KOS or metadata, in a work domain or work setting.

<table>
<thead>
<tr>
<th>Table 2: Criteria for inclusion of articles in literature review</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Articles in core library and information science journals</td>
</tr>
<tr>
<td>(Tuomaala et al. 2014), in addition to the journal Knowledge Organization, indexed in Web of Science.</td>
</tr>
<tr>
<td>2. Articles in the journal Cataloging &amp; Classification Quarterly,</td>
</tr>
<tr>
<td>indexed in Library and Information Science Source</td>
</tr>
<tr>
<td>3. Articles published in year 2000-2019</td>
</tr>
<tr>
<td>4. Articles written in English</td>
</tr>
<tr>
<td>5. Articles reporting on design of KOS or metadata</td>
</tr>
<tr>
<td>o Work domain oriented approaches included</td>
</tr>
</tbody>
</table>

Last, I read the 18 included articles and interpreted qualitatively the warrants used for design. When interpreting the warrants I have taken into consideration both research objects, that is what or who have been included in the design process, and research method or approach, that is how the design process has been carried out. In particular
the methods sections gave information on these issues. The warrants illuminate where and how the concepts and terms in the designed systems are grounded.

4. Results of the study

The 18 included articles are presented in tables 3-5 and organised chronologically within each table. I interpreted information sources’ warrant, task performers’ warrant, and work context warrant as the main warrants in the articles. The tables show the year the articles were published, the author(s) of the articles, the type of KOS designed, and the work domain connected to the design.

4.1. Design based on information sources

Eight of the articles have based the design mainly on information sources. It seems that information sources' warrant more often leads to faceted classification or taxonomies in these articles as seen from Table 3.

<table>
<thead>
<tr>
<th>Published year</th>
<th>Author(s)</th>
<th>Type of KOS</th>
<th>Work domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>Chaudhry AS; Jiun TP</td>
<td>Taxonomy</td>
<td>Integrated museum and archives system</td>
</tr>
<tr>
<td>2005</td>
<td>Chaudhry AS; Ling GH</td>
<td>Taxonomy</td>
<td>A business consulting company</td>
</tr>
<tr>
<td>2008</td>
<td>Giess MD; Wild PJ; McMahon CA</td>
<td>Faceted classification</td>
<td>Engineering design documentation</td>
</tr>
<tr>
<td>2010</td>
<td>Menard E; Mas S; Alberts I</td>
<td>Taxonomy</td>
<td>Artefacts Canada digital collection</td>
</tr>
<tr>
<td>2010</td>
<td>Wang Z; Chaudhry AS; Khoo C</td>
<td>Taxonomy</td>
<td>A graduate department in information studies</td>
</tr>
<tr>
<td>2012</td>
<td>Doria OD</td>
<td>Faceted classification</td>
<td>A department of an industrial group</td>
</tr>
<tr>
<td>2016</td>
<td>Hubain R; Wilde M; van Hooland, S</td>
<td>Controlled vocabulary (SKOS)</td>
<td>A company within the biopharmaceutical industry</td>
</tr>
<tr>
<td>2019</td>
<td>Sarasa Cabezuelo A</td>
<td>Taxonomy</td>
<td>Mathematical logic course</td>
</tr>
</tbody>
</table>

Chaudhry and Ling (2005) used sources in the form of lists of terms seen as representative of the business consulting environment, while also keeping in mind the company’s information needs. Other types of sources used in the design process are working documents “from the viewpoint of document usage” (Doria 2012, 286), engineering documents forming “a realistic representation of the domain” (Giess, Wild, and McMahon 2008, 388), and additional sources “related to the tasks of the stakeholders” (Wang, Chaudhry, and Khoo 2010, 258). The information sources’ warrant is here connected to both work context and task performers, in that the sources should represent the context or the domain, or be related to usage or tasks.

The choice of sources to use for design of a KOS is an important issue in these articles. The conclusion of Wang and colleagues (2010, 267) was that “the DDC and the domain thesauri were far from being sufficient for the organizational taxonomy”, one of the reasons being that the organisational taxonomy depended on the activities of the
organisation, and the tasks of the stakeholders, and DDC was based on disciplines. Another feature of the design processes is the involvement of task performers through informal discussions on information needs, and an informal survey on categories (Chaudhry and Ling 2005), interviews on tasks and finding information resources, and review of the taxonomy draft (Wang et al. 2010), and observations of individual folder organisation of documents (Doria 2012).

Chaudhry and Jiun (2005) gathered concepts and terms from internal and external KOS in the cultural heritage domain, and visits to museums, in order to develop a taxonomy for a cultural heritage network. Ménard, Mas, and Alberts (2010) developed a taxonomy for Artefacts Canada by means of faceted classification. Several different sources were used: terminology sources in the museum community, and existing database fields. User testing was carried out as part of the design process, however the users were not tied to any specific organisation or work domain. Hubain, De Wilde, and van Hooland (2016) used natural language processing and machine learning to create a controlled vocabulary for a company within the biopharmaceutical industry. The sources used for collecting terms were both structured resources (two biopharmaceutical glossaries) and two textual corpora (the first consisted of standards, norms, and scientific papers, and the second was based on internal documentation). Thus, the sources were connected to the work domain in question, even though automatic methods were employed. Sarasa Cabezuelo (2019) made a taxonomy for a mathematical logic course, using an existing subject taxonomy as a base, that is the Subject taxonomy for mathematical sciences education. According to the author the added entry represents concepts within propositional logic and predicate logic. The author referred to an encyclopedia of mathematics education when discussing the design decisions, thus it seems that educational warrant is also relevant here.

4.2. Design based on task performers

Six of the articles have used mainly a task performers’ warrant. Different types of KOS were suggested in the articles: thesaurus, metadata, ontology, and taxonomy, as seen in Table 4.

<table>
<thead>
<tr>
<th>Published year</th>
<th>Author(s)</th>
<th>Type of KOS</th>
<th>Work domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>Lykke Nielsen M</td>
<td>Thesaurus</td>
<td>A company within the pharmaceutical industry</td>
</tr>
<tr>
<td>2005</td>
<td>Ferraioli L</td>
<td>Metadata</td>
<td>A health care agency</td>
</tr>
<tr>
<td>2007</td>
<td>D'Ambrosio, DM</td>
<td>Metadata</td>
<td>Systems analysts</td>
</tr>
<tr>
<td>2011</td>
<td>Paling S</td>
<td>Metadata</td>
<td>American literary community</td>
</tr>
<tr>
<td>2011</td>
<td>Pattuelli MC</td>
<td>Ontology</td>
<td>Middle and high school social studies teachers</td>
</tr>
<tr>
<td>2016</td>
<td>Lund H; Ørnager S</td>
<td>Taxonomy</td>
<td>A United Nations organisation</td>
</tr>
</tbody>
</table>

Interviews and surveys were employed to investigate task performers' views and experiences. The interviews are made from a work viewpoint. The task performers are interviewed about their information-searching behaviour (Lykke Nielsen 2001); search and use of sources (Pattuelli 2011); utilising the intranet (Lund and Ørnager 2016);
requirements expressions (D’Ambrosio 2007); potential metadata elements in the context of discovering and reading new literary work (Paling 2011); and the personal construction of classification schemes (Ferraioli 2005).

In addition to interviewing task performers, several of the articles include some type of textual sources in the design processes. As for other warrants used, Pattuelli (2011) used a collection of digital primary materials, Ferraioli (2005) used incoming documents, Lund and Ørnager (2016) used an existing taxonomy and search query log data in a card-sorting exercise, and Lykke Nielsen (2001) used written user requests as well as other sources. Search query log data and written user requests can also reflect a user warrant. Lykke Nielsen (2001) used a domain study and person-in-situation approach, and can in this respect also be an example of work context warrant. Lykke Nielsen’s conclusion was that “Knowledge and understanding about work tasks and working conditions made it possible to shape the role and topical focus of the thesaurus” (2001, 793).

4.3 Design based on work context

Four of the articles have used mainly a work context warrant. These studies are varied, and different types of KOS are designed: classification scheme, taxonomy, thesaurus, and information architecture.

<table>
<thead>
<tr>
<th>Published year</th>
<th>Author(s)</th>
<th>Type of KOS</th>
<th>Work domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>Vernitski A</td>
<td>Classification</td>
<td>Humanities scholars</td>
</tr>
<tr>
<td>2010</td>
<td>Milne C</td>
<td>Taxonomy</td>
<td>University portal</td>
</tr>
<tr>
<td>2015</td>
<td>Soglasnova L; Hanson M</td>
<td>Thesaurus</td>
<td>Non-profit community legal aid clinic</td>
</tr>
<tr>
<td>2017</td>
<td>Ruzza, M et al.</td>
<td>Information architecture</td>
<td>An Italian public organisation</td>
</tr>
</tbody>
</table>

Soglasnova and Hanson (2015) explored the techniques of cognitive work analysis to evaluate a thesaurus for the library catalogue of a community legal aid clinic. The methods included a questionnaire and user interviews, word association tests, and a hierarchy exercise. The authors analysed responses to the user interviews by means of dimensions in the cognitive work analysis framework. Cognitive work analysis is a work-centred framework for analysing information behaviour in context (Fidel and Pejtersen 2004). It was found that the work environment “is constrained by two differing models of discourse” (Soglasnova and Hanson 2015, 915) and thus the social context played a central role here. Milne (2010) used functional classification and illuminated functions, activities, tasks and sub-tasks in a taxonomy for a university intranet portal. After defining the function of quality assurance/enhancement, Milne used a higher education business classification scheme followed by analysis and discussion to identify the activities. Tasks were identified by interviewing colleagues. The terms used in the taxonomy reflect the functions, activities, tasks and sub-tasks in the organisation. This approach can be a mix of work context warrant (functions and activities) and task performers’ warrant (tasks and subtasks).

Vernitski (2006) developed a classification scheme for fiction for Humanities scholars. The author argued that intertextuality is central to scholars of fiction, and
therefore the intertextual approach was chosen as a basis for the classification scheme. In the development of the scheme the author created categories and sub-categories of types of intertextual links. In this manner the basis for the scheme is both the needs of task performers (humanities scholars), and the scholarly domain itself (humanities and intertextuality). A scholarly domain can constitute a scientific or academic warrant, but it can also be interpreted as a context for scholars, as a work context warrant. Ruzza and colleagues (2017) developed a model for the design of information architecture for websites of large organisations. The model was both developed and tested through action research with an Italian public organisation. The methods included an online survey, content mapping and analysis of the website, review of literature, and short interviews with web content managers. The basis for the information architecture is the organisation’s need for a well functioning website, that is a type of work context warrant.

4.4 Model of warrants
As seen from the results, the three main warrants were often used together for the design of KOS. The information sources often originated in an organisational or work context, and the task performers were often part of an organisation or a domain. The use of information sources was often complemented with interviews with task performers. This combination of warrants is illustrated in Figure 1.

4. Discussion
The concept of warrant is interesting and useful for enterprise search as seen from this study. Warrant helps to focus on the methodological level in the design process. It provides a more specific level of detail when discussing design of metadata for enterprise search. It illuminates how and where the terms and structure in KOS and metadata are grounded. Warrant is able to add to the more general division between manual and automatic methods for the design of KOS for enterprise search (Cleverley and Burnett 2015). Information sources were used in several articles in this literature review to elicit terms and relations, both for manual and automatic design. At the same time the information sources used in the studies reflect the work domains in question. The terms gathered are not a priori decided on, they are connected to a context. Interviews with task performers provided insight into how they searched for and used information as well as what metadata elements were deemed useful. The role of work context is interesting. Sometimes as a background for design, and sometimes as a direct
source for terms and relations between terms. For example in functional classification the context is more directly reflected in the terms (functions, activities, and tasks).

Several different types of KOS or metadata were described in the articles in this literature review: taxonomies, classifications, thesauri, ontologies, metadata, and vocabularies. The results do not suggest that any particular warrant leads to any specific type of KOS, or that any warrant is more useful than another. What the results do say is that a mix of different warrants are often used together in a study. This suggests that it is worth considering both information sources, task performers, and work context in the design of KOS for enterprise search.

Whereas this study argues that the concept of warrant is useful for the design of KOS, other concepts and models seem necessary for the implementation and evaluation of metadata in enterprise search. One such model is the theoretical model in Cleverley and Burnett (2015). This model shows on the one hand the KOS (manually and/or automatically developed). On the other hand is the implementation of the KOS in enterprise search (manual and/or automatic classification or categorisation). One possibility in the model is also automatic organisation of information without using any KOS. The model also includes two types of search: known item search and exploratory search, and suggests for example that known item search is supported by manually developed taxonomies (Cleverley and Burnett 2015, 439). Another model relevant for enterprise search is the evaluation framework of user interaction with metadata surrogates (Balatsoukas, Morris, and O’Brien 2009). This model suggests different possibilities for evaluating user interaction with metadata surrogates in a search results interface.

6. Conclusion

This study has analysed journal articles that design metadata or KOS in the domain of work. The results indicate that information sources’ warrant, task performers’ warrant, and work context warrant are the main warrants used in design. The results are interesting for the design of metadata for enterprise search, because these warrants can provide help in choosing different methodological perspectives when designing metadata. Other concepts and frameworks are relevant for implementation and evaluation of metadata in enterprise search.

One limitation of the study is that the procedure used for searching and selecting articles might have missed potential interesting articles. Other journals than the selected library and information science journals could have provided a different picture. More weight on automatic methods, and conference proceedings, could have supplemented the results.

References


http://informationr.net/ir/10-1/paper210.html


Abstract:
This article aims to present two different approaches related to the study of types of documents. Our research problem is to understand to what extent the identification of the document type reveals the persistent representation of the archival document. We will present two different approaches related to the study of types of documents: one about the types of documents as a social product, which represents social practices and recognizes that social practices control not only the behavior of the individual, but also the structure and form of the document. From this perspective we will analyze documents from personal collections created in the 19th century. The other approach focuses on the organizational environment of a federal public agency and how document types can be used as an instrument of transparency in a records management system. The research will focus on digital documents created in an electronic system. The study shows that document types also provide the basis for classification systems that are the foundation for acquisition, organization, description and evaluation. Being the closest representation of the acts, the records are taken as evidence, that is, they reveal the persistent representation of the archival document, contributing to the organization of knowledge about the collections, their producers and about society in a given historical period.

1.0 Introduction
The approach to the organization of knowledge has different perspectives. One of its main objectives is permitting knowledge to move from a purely theoretical nature towards the creation of different processes and activities that serve to disseminate information. This statement finds support in the statement of Guimarães (2017, 92), for which, among the dimensions of Knowledge Organization (KO) - epistemological, cultural and applied - the latter would translate into:

“manage and identify large pieces of information which leads on the other hand to the need to develop increasingly friendly tools, making the organization and knowledge representation processes less and less artificial and closer to the user's daily life” (Guimarães 2017, 92).

In other words, the author points out that the organization of knowledge permeates different actions that, in the end, contribute to the KO providing users with relevant and appropriate content for their needs. In this sense, we can affirm that in the field of archives the study of document types is one of those contributions.

In the context of records management the study of types of documents has been used more frequently as a tool for appraisal, classification and destination. This practice echoed the Madrid (Spain) school and has spread throughout Brazil. However, for the past ten years, the method has also been used in Brazil to understand personal archives. From these two applications of the document type analysis method the research discusses to what extent the identification of the type reveals the persistent representation of the archival document.

One scenario seeks to understand how the documents produced by the creator of a file in a given historical period can describe the connection between the individual and
society, while the other scenario focuses on the organizational environment of a public agency and how the identification of types of documents can be used as an instrument of transparency in an information management system. Both situations place records and files at the epicenter of the discussion on representation and therefore contribute to the organization of knowledge in a given social and historical context.

We consider that the type of document and its concept offer the means to understand the context of production, revealing its archival link and equally making access more instrumental. This occurs in so far as the archival document under any and all circumstances is a social product since it represents social practices. This study identifies the semantic and cognitive descriptions used to name types of documents and recognizes the connection with the action that is the source of the record. We emphasize that the recognition of social practices as tacit rules that control not only the individual's behavior, but also the structure and form of the documents, provides the key to understanding the context of the archive and improves the use of the archives.

Geoffrey Yeo points to the concept of record representation, defining that when we speak of a record we speak of “persistent representations of activities, created by participants or observers of those activities or by their authorized proxies” (Yeo 2007). This concept was used as part of the theoretical foundation of the research.

In the context of records management, the identification of types of documents and the functions they represent provide the basis for the acquisition of knowledge about archival production, about the archival link that connects the producer and the documents, and about the mission and functions of the institution producing the archive. This knowledge will be used to guide access, management and appraisal policies, and finally, it will make documents accessible for accountability and active transparency.

In the end, the article will show that the use of both approaches enriches the process of understanding archives, provides a consistent production of knowledge about records and offers a perspective that meets society's wishes.

2.0 Archival representation and the archival document

The understanding of the adopted archival representation is based on ideas of Yakel (2005) and Yeo (2007). According to Yakel:

“the term archival representation will be used for the archival function commonly and variously identified as arrangement and description, processing and occasionally archival cataloging. The term “archival representation” more precisely captures the actual work of the archivist in (re) ordering, interpreting, creating surrogates, and designing architectures for representational systems that contain those surrogates to stand in for or represent actual archival materials” (Yakel 2003, 2).

For the author, it is important to revisit the representations elaborated in regard to the collections, since the cultural context of the production of these representations is dynamic. This action of revisiting them makes it possible to complete them and even correct them. The author also proposes that the production process of arrangement and description be continuous and fluid.

“The archival representation, then, demonstrate not only the evolving physical collections and intellectual understandings of collections, but also changing perspectives on collection arrangement, descriptions, and management. Each successive representation and representational system builds on its predecessors, recovering what was judged valuable in a given temporal and cultural context, incorporating or discarding what was deemed essential or not respectively” (Yakel 2003, 24).

For the author:
“Archival representations speak not only about the collections for which they act as surrogates, but also archival practice and archivists” (Yakel 2003, 25).

In Yeo's view, archival representation is intrinsically related to the archival document. The author, when approaching what he understands by archival representation, revisits the conceptions attributed to the archival document (evidence and information). He then defines the representation of the document as follows:

“To differentiate records from other kinds of representation, records can be characterized as persistent representations of activities, created by participants or observers of those activities or by their authorized proxies” (Yeo 2007, 337).

Yeo also has three attributes to characterize the persistent representation of archival documents: the document has the ability to represent the activity that originated it over time and to allow its sharing; it represents the most diverse activities of actions carried out by individuals, not just the legal, economic and evidential transactions; and is produced by observers of, or participants in, the represented activity. Yet, we identified two characteristics that still lack further exploration in theoretical discussions regarding this area but that are fundamental for the preservation of documents: 1. they are also persistent in their ability to be a vehicle of information and 2. the stability of instrumental content supplies different interpretations of the same event or action.

In our view, the ability to represent the activity of the individual or institution that generated the document is what guarantees the different uses and functions that is ascribed to it. Thomassen (2001, 375-376) in his article "A first introduction to archival science" points out the following functions of the archive document: "memory of individuals, of organizations and of society, support operational management, agents of accountability, evidence and cultural and historical function". Society's use of documents and their subsequent importance touch, among other things, evidence, information, memory triggers, accountability, justice, identity, power, citizenship. Each of these purposes of the document's use also reveals society's own understanding of its instrumentalization and the socio-political context in which this occurs.

The activity of identification of the document type reveals, therefore, the genesis of its production and in this way we demonstrate the persistent representation that each archival document has and the vestige of its future use by society. In addition, the emphasis on the activity of production also indicates the relevant archival context. In this sense, the archivist is at the epicenter of archival representation in terms of the archival document itself, and also when the typological identification is performed in the application of the above formula by the archivist, thus ratifying what Yakel and Yeo understand by this type of representation.

3.0 The document type applied in the organizational environment and in personal files

This section will present two approaches to the application of the document type identification methodology from two perspectives: that of personal archives and that of institutional archives. With regard to personal records we have applied the following definition by Oliveira (2012):

“We understand "personal archives" as a set of documents produced or received and maintained by an individual throughout his life as a result of his activities and social functions. These documents, in any form and on any medium, represent the life of the holder and their personal or business networks. They also represent intimate and personal aspects, as well as the creator's deeds, etc. They are obvious records of the creator's broad role in society” (Oliveira 2012, 33).
Carmargo affirms that personal archives are indeed archives and must be anchored in the context of their production so that they do not lose the representative aspect that makes them unique and singularizes them. The author states that:

“In order to guarantee support for the different possibilities of interpretation that the reading of their content can raise over time, personal archives must be treated as proper archives, that is, they must be anchored to the context in which they were produced. When this relationship is subverted, that is, when the potential for use, taken in its inexhaustible and imponderable magnitude, enters as a component of the treatment of the archives, replacing actions that justified their production, the documents lose the representative aspect that makes them unique” (Camargo 2009, 36).

Institutional archives are the set of documents received and produced by an organization when carrying out its activities. This understanding underpins the very concept of these archival documents.

Before presenting our understanding of document types we will present the kind of document which is the configuration that a document assumes according to the disposition and nature of the information contained therein (Camargo and Bellotto 1996). The concept of document type defended by Antonia Heredia Herrera consists of: “The structure, the material form in which the content is materialized” (Herrera 2007). Bellotto defines it as follows: "document type is the documentary form assumed according to the activity that generated it, its fixation is complex and punctual” (Belloto 2002, 91). From the definitions above, when identifying a document type we apply the following formula: kind plus activity.

The applied methodology broadens the study in the field of personal archives since the individual in his legacy of documents gathers records relating to his intimate life and to a life subject to a society structured by rules and legislation. This breadth enriches the study of the personal archive and enables the production of knowledge regarding a historical period. The analysis is based on a methodology developed and tested in continuous research projects since 2003 that were developed and coordinated by Oliveira using a typological identification form that contained 10 fields: kind, activity, type, date, location, recipient, activity (-ies) / reason that originated the document, specific typical characteristics and standard elements. The filling instructions are detailed below:

“1 - The Kind field aims to demonstrate the configuration that the document assumes according to the nature of its information. 2 - The Activity field seeks to identify the objective that drove the production of the document. 3 - The Type field seeks to define the configuration that the documentary kind assumes according to the activity that generated it (kind + activity). 4 - The Date field marks the period in which this document was produced and, consequently, in which social posture code it fits. 5 and 6 - The Location and Author fields help us to locate these documents. 7 - The Recipient field allows us to identify whether the document was produced for personal or professional purposes and for whom it was produced. 8 - The Activity (s) / reason field that originated the document explains the circumstance in which it was produced. 9 - The Specific Typical Characteristics field helps us to describe the characteristics present in the document and which relate it to the event to which it belongs, either in written form or through its symbols. 10 - The Standard Element field allows us to find the constant elements found in the documentary set referring to the specific event and which give rise to their typology” (Oliveira and Penna 2013, 483).

As we mentioned, the research makes use of documents preserved in personal and institutional archives. We will use a sample of documents from each context (the management of archives and personal collections and documents produced by a public institution) since there are more than 800 types of document defined so far. It should be noted that the work of identification continues and this number may be changed. In this sense,
we work with 30% of the group. The study provides a comparison of the two systems and analyzes the need to keep all the acts represented over time.

When we preserve personal archives we generally acquire documents that reflect the professional activities of the owners of the archives and the legacy that proves the notoriety of the individual or individuals who originated the documents. Records of intimate or ordinary life in general are not preserved. The set of archives with which we work has this ordinary and human dimension that allows us to get to know more closely the social and cultural context in which the producer of the archive and the individuals who were part of their network of relationships lived. The analysis methodology involved researching the types of documents identified, their concepts and their relationship with the archive and with the historical context. For the purposes of illustration we have chosen two types of documents: a rural property management report and a slave birth certificate.

Rural property management report: Report on decision-making activities for the development of activities on a rural property.

![Rural Property Management Report](http://www.memoriaes-cravidao.rb.gov.br/exposicoes/Catalogo1/#p=19)

We observed that the form of the document could induce, in a faster analysis, its identification as a letter, which among its most striking characteristics is that it is a written message sent from one individual to another, containing place, date, name of the recipient and signatory, and if an official message it would typically contain number, letterhead and address. The form of treatment depends on the nature of the letter, whether personal or official, and the degree of intimacy between the interlocutors. This should be noted at the beginning, when the recipient is mentioned, and at the end of the farewell when the sender signs. In private life identification is often by nicknames.

The report (the documentary form) has some of the characteristics of a letter, so in a hasty analysis it can be mistakenly identified. The report type is understood to be the recorded narrative or exhibition about a certain event or fact and its content must be about expenses, activities and commercial transactions. Its frequency is variable and there is usually a subordinate relationship between the author and the recipient that must be identified as well as their position or role. The motivation for producing the report
must be clear and have a date. In the end you may or may not have a recommendation. The understanding of the document's function is that it will give the precise indication of the type and with this information and construction of a concept that standardizes its use by archivists and by users it is possible to consolidate knowledge about the archive, the producer and its functions, persistently loading this metadata over time for access and retrieval.

The second example is produced in the family environment but follows legal guidelines so that it can be accepted in society. The declaration (documentary form) is considered as a written statement about a fact, event or right. Usually, it has the title of Declaration, the name and identification/title of the declarant, what is being declared, place, date (of the fact and of the document) and signature of who is declaring the fact.

Figure 2: Slave birth certificate CFBO SFJ DAAFJ 244 (http://www.memoriaescravodao.rb.gov.br/exposicoes/Catalogo1/#p=40)

The declaration of the birth of a slave (the type) was mandatory since it was used to notify the registration officer of the birth of a slave and then the slave was registered with the municipal body. Its structure was defined by Decree no. 4835 of December 1, 1871, which approved the regulation for the special registration of slaves and free children of slave women. The declaration should contain the name of the owner, the day of the child's birth, the mother's name, the child's sex, the child's first name and color. In addition, there should be the mother's two registrations, the one made by the owner and that of the municipality, as well as her activity. ¹

¹ In Brazil, on September 28, 1871, Law No. 2,040, known as the “Free Womb Law”, was signed. The Law declared the children of slave women born in Brazil to be free. However, the child stayed until the age of eight with the master of his mother.
The precise definition of this document type reflects the practice governed by legislation to control the birth of children of slaves and the state’s information on slave activity, thus permitting a thorough study of the period and the relationship between the state, slave masters and the slave and her children in a period of transition.

At this point, we will detail the second approach related to the study of the document type, applying it in an organizational environment. For the purpose of clarifying the origin of this approach it is necessary to know that the Federal Government of Brazil, in 2015, created a legal instrument for the establishment of the national electronic process, Decree No. 8,539, of October 8, 2015 (Brasil 2015) which provides for the use of electronic means to carry out the administrative process within the bodies and entities of the direct, autarchic and foundational federal public administration.

This decree determined that public agencies of the executive branch would have two years to implement software designed to produce and process administrative processes electronically. The Brazilian government chose the Electronic Information System (Sistema Eletrônico de Informação) SEI software. As a result of its implementation the procedural process was speedy and the acts were transparent.

In this article, we will deal with the results of the implantation of the SEI at Fundação Casa de Rui Barbosa (FCRB) specifically about the reflections around the identification of the document type as an instrument of access and transparency. FCRB fully implemented the software in 2017 and one of the activities that preceded the implementation was the identification of the document types that integrated the administrative processes and the insertion of the software layout. These actions were developed by the institution's team of archivists. The methodology for identifying document types proved to be crucial for the reproduction of the business flow and the team followed the same method adopted in the previously mentioned personal archives.

This project was entitled PenSei2 digital and was carried out within the scope of the agency's Document Management Program that strives for archival classification, for maintaining the filing of administrative proceedings, for the typological identification of documents and for the archival control of documents produced and processed within the software (Panisset and Jaccoud 2019).

The process of identifying document types was carried out before the software was implemented and emphasized the application of the first three fields defined by Oliveira and Penna (2013), as we were creating document templates that had not yet been produced which meant that consequently it would not be possible to identify the date and circumstance of the document production. Brazilian legislation that regulates administrative procedures for operation, purchase and services in the Brazilian Public Administration was also used for decision making regarding the identification of the document type, in addition to manuals on administrative procedures of public agencies and specialized dictionaries. For the purpose of demonstrating that the document type is the closest archival representation of the act and the reason for the production of the document, we will use the archival collection acquisition activity.

---

2 The project name is a combination of the acronyms of the National Electronic Process (Processo Eletrônico Nacional) - PEN and Electronic Information System (Sistema Eletrônico de Informação) - SEI. SEI was developed by the Federal Regional Court of the 4th Region. The software is an administrative process management system, composed of functionalities for administrative optimization.
The regulation of this activity provides that the acquisition of collections should be formalized in an administrative process that contains a list of documents, the most important being the donation contract, a document that provides for the common agreement between the parties involved (the donor and Fundação Casa de Rui Barbosa) that demonstrates the willingness to donate, the identification of the documents to be donated, the conditions of access, reproduction and preservation. In addition to the contract, the process comprised: the request to acquire a personal archive; the decision-making order; donor identification and civil registration documents; the identification list of the archival collection; the declaration of sole heir; the checklist of administrative acts and documents, legal opinion and donation contract (with or without copyright). Below is an image of the set of documents that comprise a collection archiving process and the acquisition contract (where the personal information of the donor has been deleted).

Figure 3: Pensei’s homepage/ SOURCE: SEI
Currently, the software has 677 document types, 182 with models (documents produced within the software) and 495 without models (documents that are digitized and captured into the system). The study and use of the identification and definition of document types in Document Management promotes the appropriate choice of the archival document to be produced and further enhances access to this document both to support management in the decision-making process, as well as to serve society and give transparency to administrative acts promoted by Fundação Casa de Rui Barbosa.

4.0 Final Considerations

Both approaches adopted in this article point out the relevance of the typological study to obtain the improvement of processes that are crucial to public administration and society. In response to our research question, we understand that the identification of the document type is essential for the persistent representation of the archival document, which may have different purposes, since it records the genesis of the documents and links the document to the processes in which the producer of the archive participates and unveils the context in which the actions took place, as well as their actors and developments. It is the closest to the functionality and dynamics of fact and consequence, and it is undoubtedly the mechanism that sustains the typical persistence of archival documents.

The study pointed out that the types of documents provide the basis for the classification systems, which are the fundamentals of acquisition, organization, description and evaluation, and an important axis for the evaluation analysis, since it brings out the essence of the represented action. Being the closest representation of the acts, the records are taken as evidence. But they can be more than that, as they are social products and a trigger for our individual or collective memory.

We understand, with the application of the methodology in the different circumstances, that the document type is a social product and the persistent representation of
both the activity producing the document and the document itself. Therefore, aligning the concepts of knowledge organization and document types, we realize that this process of archival nature meets the demands of the KO, recognizing the importance of its use and application in different contexts that require theoretical and practical understanding of the role of the archive, the archivist and, in particular, archivology in the promotion and diffusion of organizational or social knowledge.

References
Organizing Performing Arts Records of Korean Traditional Music as Linked Open Data

Abstract:
This study analyzed the processes for publishing and providing archival records of Korean music performance materials in the form of linked open data. The subject of the analysis was records of performing arts materials in the Gugak Archive, which manages and services Korean traditional music performance materials. In the analysis, the research team considered if the results applied to performing arts fields and reflected the characteristics of Korean traditional musical performances and if the records could be published as linked data. The criteria for linked data was set by identifying the resource, designing the data structure, and linking related external data necessary for structuring semantic web data. The study’s results could be used in collaboration with cooperate with related organizations that provide Korean traditional performing arts materials or in the improvement of Korean traditional music archives through linked data.

1.0 Overview
This research provides analysis and suggestions in building linked semantic data based on the records of the Gugak Archive of the National Gugak Center (NGC). Creating semantic web data enables the information of performing arts stored in a separate data silo to be searchable and accessible in a web environment. In general, separated data stored in a database are processed according to the intention of a service designer. Therefore, data use is limited within the given scope and structure. In addition, the amount of data that can be downloaded or browsed is also limited to avoid website traffic problems or slowdowns. In contrast, linked data expands the scope of sharing and processing raw data in the semantic web environment. The research team analyzed the performing arts records, focusing on the identification of the data unit, modeling of the data structure, and connection of relevant information (Park et al. 2019).

The characteristics of performing arts materials and Korean traditional music performances were also considered because the event performance contained in the material should be expressed in the archival records, and related performance materials should be grouped together. Furthermore, improving the format and content of data will also increase the quality and usability of the data. The team analyzed the classification and cataloging information of gugak performing materials in the Gugak Archive and also examined how archival records of gugak performing materials are modeled based on the Functional Requirements for Bibliographic Records – Object Oriented (FRBRoo).
2.0 Gugak Archive of the National Gugak Center

“Gugak” refers to Korean traditional music that involves performing arts, including modern contemporary gugak. The Gugak Archive, an organization affiliated with the NGC, collects and manages the data of gugak performances. The NGC, established in 1951, officially launched the Gugak Archive in 2007 for the collection, management, preservation, and service of Korean traditional music resources. Since then, the archive has collected Korean traditional materials and established a classification system and descriptive elements for cataloging (Kweon 2016; NGC 2019). In addition to managing gugak performing resources, the Gugak Archive also provides materials of gugak music performances through its website and external portal search site.

3.0 Characteristics of gugak performances and their materials

The NGC develops gugak performance themes, such as “Saturday Gugak Concert,” as a part of its brand and conducts them throughout the year. At times, the center also holds special performances and events, such as the “New Year's Gugak Concert,” which are only performed for one day. In addition to the NGC’s performance materials, the Gugak Archive also publishes or collects information materials about gugak performances, including valuable donated materials. Therefore, the Gugak Archive’s performance materials have multiple types and diverse characteristics. The following is the research team’s analysis of the characteristics of gugak performance materials:

- There are more series of performances by themes than individually authored performances. E.g., “Saturday Performance of Korean Music & Dance,” which takes place every Saturday.
- One performance consists of many individual gugak works. Individual performance planning is composed of various musical works. E.g., “Saturday Performance of Korean Music & Dance” (May 21) was composed of individual works such as “Daechwita,” “Gayageum Sanjo,” “Ajaeng Sanjo,” and “Sinawi” performances.
- Various types of archival materials related to performance planning, such as brochures, posters, invitations, and scripts, are produced before a performance.
- Different types of performance materials, such as audio and video recordings and photos, are provided after a performance.

4.0 Organizing the gugak performing arts materials

The Gugak Archive created an internal guideline for classifying and describing gugak materials (Roh 2017). They developed “Gugak classification scheme” based on KDC (Korean Decimal Classification), which is connected to “679, traditional Korean music” class. They also created a gugak glossary and a gugak thesaurus to utilize when indexing gugak materials. Currently, the Gugak Archive assigns a class name and keywords to each item and its components (clips). Description rules follow an internal metadata guideline created based on the General International Standard Archival Description [ISAD(G)] and Dublin Core Metadata Element Set. Under this guideline, gugak information is divided into three units: performance (collections), performance materials (items), components of performance materials (clips). All performance materials can be
grouped per performance as a subcomponent. Then, they are further divided into sub-components with classification, index, and descriptive information.

5.0 Characteristics of searching and browsing the online catalog

The Gugak Archive provides a searching and browsing service for its records on the website. However, there are several semantic issues such as below:

1) Linking information between performances and their materials: Performance materials should be grouped per performance so that users can browse them. Currently, the Gugak Archive grouped performance materials by performance internally to manage them, but the grouping information is not accessible on the website. Moreover, information between the NGC’s performance database and the Gugak Archive’s records is not connected directly.

2) Identifying and accessing the information of gugak performance materials using Uniform Resource Identifier (URI): Currently, there is no URI specified for consistent access to the gugak archival information. Users are able to access the information of the gugak materials by copying their Uniform Resource Locator (URL) listed on a web page. However, the URLs for each webpage could be changed at website renewals. Moreover, the hierarchical relationship between performances and their materials is ambiguous in the URL structure. The hierarchical structure of performance materials is composed of Collection, Series, Folder, Item, and Clip. However, the scope of Collection or Series is too wide, or the Folder and Item hierarchies are often set up identically. Hence, it is necessary to adjust the conceptual level of each hierarchy clearly.

3) Converting descriptive information in text format into structured access points: To build rich semantic linkage information, metadata values should be controlled by access points rather than descriptive values. However, important information, such as gugak performance programs or lists of performers, is entered only as text. The Gugak Archive also carefully describes the abstract characteristics of performances (i.e., genre of the performance work) based on the materials and the physical characteristics of performance materials, such as format or size. The archive also provides the classification information and search keywords. However, the browsing menu of the Gugak Archive website displays information about the performance item and individual component in one depth. Moreover, the information about the individual performance work, physical type of the item, or component are mixed in a keyword cluster. In this case, descriptive elements and access points should be clearly specified to enable users to search and utilize the materials at various intellectual and physical levels. The link information between individual access points should also be established.

6.0 Identifying and accessing gugak performances and their materials

Based on the analysis above, the research team suggested a structure that improves the identification and accessibility of gugak materials in a semantic web environment; and applied it to the leading performance of the NGC, “Saturday Performance of Korean Music & Dance” series, by demonstration.
1) Analysis of “Saturday Performance of Korean Music & Dance” and its materials:
   a) Characteristics of “Saturday Performance of Korean Music & Dance:” This year-round performance has been around for more than 30 years in the National Gugak Center. It embraces various genres: court music, folk music, instrumental music, and vocal music. Its name has been changed a few times. The performance is usually planned on a yearly basis, and each month’s performance is composed of subcomponents such as “type A” and “type B.” However, even if “type A” is played repeatedly, various individual musical work play every week.
   b) Characteristics of the Saturday performance planning materials: Various materials are produced, such as posters, brochures, and invitations. Posters introduce performances on a yearly basis. Brochures promote performances per quarter. The scope and depth of the information included differ according to the time range introduced in the materials.
   c) Characteristics of the Saturday performance’s records (Post-performance materials): Live video or audio recordings and photos are created. For video recordings, there are original (filmed by a camera) and master (edited) versions with occasional additional commentaries. Most of the data are created on the performance day, and some of them have video or audio clips as subcomponents. Specific information is also written for individual musical works, musicians, and performing artists.

2) Requirements for assigning URIs and linking related entities between archival records:
   a) URIs for the performance plan
      • URIs for plan: Specify URIs to the performance planning per corresponding year and connect each URI to tracking the change of the title of the performance: 1998 Saturday Gugak Stage → 2001 Saturday Year-Round Gugak Performance → 2002 Saturday Gugak Stage → 2005 Saturday Master Artists Performance → 2006 Saturday Year-Round Performance → 2009 Saturday Year-Round Performance → 2010 Saturday Performance of Korean Music & Dance (up to date)
      • URIs for subplan: Indicate the sub-title of the performance planning, which always follows, to the URI as a subcomponent. E.g., 2011 Saturday Performance of Korean Music & Dance: Intangible Cultural Heritage of Humanity → http://domain/performance_plan_id/sub_plan_id/
   b) URIs for individual performance items and their components
      • URIs for the performance: Specify URIs to the individual performance and link them to the performance plan
      • URIs for performance subcomponents: Subcomponents should be cited, such as video or audio clips, by extending individual performance URIs. E.g., 2011 Saturday Performance of Korean Music & Dance: Intangible Cultural Heritage of Humanity [May 21] → http://domain/performance_plan_id/sub_plan_id/performance_id
   c) URIs for individual musical work within individual performances
• URIs for related individual work: Individual performance (performed on May 21) within a performance plan (Saturday Performance of Korean Music & Dance: Type “C”) includes several individual musical works (e.g., Daechwita, Sanjo). Identify individual musical work and assign URI. E.g., Daechwita → http://domain/gugak_work_id

• External links for individual musical work: Individual musical work can be efficiently utilized if linked to external information such as related persons, locations, re-sources. E.g., URI for Daechwita (http://domain/gugak_work_id) can be connected to URI for subject heading “Daechwita” of National Library of Korea, <http://lod.nl.go.kr/resource/KSH00101678> (National Library of Korea 2019)

d) Relation between performance and related materials
• Connect performance materials to the performance plan and individual performance information.
• Many individual performances can be connected for a brochure, “2011 Saturday Performance of Korean Music & Dance [January 8 – December 24]”
• Posters can cover a broad range of time periods, but also specify a specific performance such as “2011 Saturday Performance of Korean Music & Dance: Type A [January 8]”
• Live video clips can be used when numerous materials are connected to a single performance such as “2011 Saturday Performance of Korean Music & Dance: Intangible Cultural Heritage of Humanity [September 24] Original ①.”

e) Constructing data about persons, organizations, locations related to the performances and their materials, and specifying URIs
• Performance materials can be related to abstract objects or events, such as the performance work, performance plan, and the performance itself, besides physical properties per media.
• Information on organizations and locations (venues) can be connected to the data extracted from the Korea Arts Management Service.
• Pioneers of gugak music or musicians/performing artists can be found in the Encyclopedia of Korean Culture.

7.0 Considerations for Modeling Abstract Concepts Using FRBRoo
In adapting FRBRoo, the Gugak Archival records, which are currently organized based on physical performance materials, “manifestation” or “item” in FRBR terms, should be grouped with abstract concepts of performance work, performance plan, and performance. For this, it is suggested to apply FRBRoo, an associated model of CIDOC CRM, for modeling the information of performing arts and their materials (Bountouri and Gergatsoulis 2011; Park 2018; Park 2019). However, the analysis of the current state of the Gugak Archive records shows that the following matters should be considered for the application of FRBRoo.
1) Units of *F20 Performance Work*
   a) *F20 Performance Work* is a subclass of *F16 Container Work*, which is either “sets of concepts” or “series of like performances.” With this, *F20 Performance Work* is differentiated from the *Work* entity of the original FRBR model and is also different from the *F1 Work* of FRBRoo that is mapped to the FRBR’s *Work* entity.

   b) In the Gugak Archive, Series and Folder units have similar concepts as *F20 Performance Work*. The Series is a continually repeating performance brand, such as “Saturday Performance of Korean Music & Dance,” which was adopted as an example in the main text. Meanwhile, Folder can be regarded as “2011 Saturday Performance of Korean Music & Dance: Intangible Cultural Heritage of Humanity [May 21],” a specific performance belonging to the “Saturday Performance of Korean Music & Dance” series. The sum of the artistic content and stage management contained in the Series and Folder unit of the Gugak Archive can be regarded as *F20 Performance Work*.

   c) In the case of Gugak, however, the types of Series and Folders are not simple. The “Saturday Performance of Korean Music & Dance” series itself is planned yearly but is transformed into various subseries within a year. They are listed as “Style Ga” and “Style Na” to change the content of the performance or specify themes, such as the “Master Program” and “Intangible Cultural Heritage of Humanity.” This characteristic of the Korean traditional gugak performance makes it ambiguous which field the *F20 Performance Work* should be associated with.

2) Scope of *F25 Performance Plan*
   a) *F25 Performance Plan*, along with the *F20 Performance Work*, is an important means of modeling performance materials in FRBRoo. As *F20 Performance Work* is a subclass of *F16 Container Work*, the expressions associated with *F20 Performance Work* have complex properties as *F25 Performance Plan*. In performances, it is a composite form of individual expressions, such as stage scripts, background music, and performance program materials. As revealed in FRBRoo, “[performance plan] may be more or less elaborate and may even foresee just improvisation.” (Mapping with actual data is difficult.)

   b) *F25 Performance Plan* is a composite version of many individual expressions, and there are diverse cases of individual expressions belonging to a subcategory. For example, the director’s stage instructions are like texts; background music is like sound, and stage setting drawings are like images and maps. Therefore, there are various descriptive elements that are necessary for the expressions.

   c) Actual performers and record managers sometimes have a different point of view from that of FRBRoo’s class definition. For example, *F25 Performance Plan* candidates belonging to one *F20 Performance Work* in the Gugak Archive include a performance poster, a performance promotional placard, and a performance program. In addition, recordings and photographs of performances are also regarded as planning materials related to
the performances. However, in FRBRoo, promotional materials are modeled and linked to $F1$ Work, the related external class. Moreover, recording materials belong to a separate expression or manifestation within the $F21$ Recording Work associated with the $F29$ Recording Event.

3) Information Related to $F31$ Performance and $F20$ Performance Work

a) FRBRoo is influenced by CIDOC CRM, and core classes are linked around event-based classes. Therefore, the temporal class, $F31$ Performance class, also plays an important role in linking performance works with related works. However, it is not easy to obtain data related to the performance itself from the archival records of performance materials. The name of the performance, the date and time of the performance, and the location of the performance are important. Most of this information is contained in the performance poster and the individual performance program. In the case of performance recording materials or photographs, it is difficult to extract performance information from the materials unless separate metadata has been contained.

b) Because of the nature of performance works, the fact that the same or similar content is repeated many times makes it difficult to confirm $F31$ Performance information. If performances are run throughout the year or run on weekends for a month, performance materials are typically produced from a few representative performances. Therefore, it is often difficult to determine which performances a particular recording or photo belongs to among individual performances that were carried out many times.

4) Information Related to $F29$ Recording Event and $F21$ Recording Work

a) FRBRoo presents not only $F31$ Performance class but also $F29$ Recording Event as an important event class. Unlike other art genres, performances disappear as soon as they are performed. Therefore, the documents containing the live performances can preserve the performances and become a new work that reinterprets the performances (Lee 2018). Therefore, in FRBRoo, when an event called “performance” meets with an event called “recording,” performance works and recording works are linked with each other. However, it is difficult to find the proper metadata to describe the recording as an event in the archival records because those who recorded the performance are difficult to identify, while the director or actors are likely identifiable in the performance materials.

b) In the case of the photograph of the performance, the photographer can often be found more easily than an individual who recorded audio or videos. This is the case when a famous artist leaves a photograph of a specific performance and distributes it as his/her work of art. However, there are few cases where information, such as photographers, shooting dates, and shooting equipment is obtained and provided with the photographs.

5) Relationship Between $F20$ Performance Work and $F1$ Work

a) Performing works with the characteristics of comprehensive art are $F16$ Container Works themselves, which means that one $F20$ Performance Work is associated with several $F1$ Work. Moreover, performing arts can
be said to have abundant external links such as information about directors, actors, original dramas, etc. However, it is not easy to identify abundant external link information directly from specific performance materials. Although performances and criticisms may refer to the original plays, motives, and specific events and music, it is difficult to simultaneously organize the related individual works when recording performance materials in the archives.

b) The archival records are commonly built around the records of the collection materials. However, $F_1$ Work related to $F_20$ Performance Work is more often collected by other external institutions rather than being collected with performance materials by relevant institutions. For example, there is the Jongmyo Jeryeak (ritual music), which is a UNESCO Intangible Cultural Heritage and the first National Intangible Cultural Heritage of Korea. The Gugak Archive contains diverse performance materials related to Jongmyo Jeryeak. However, as most of the Jongmyo Jeryeak is performed as part of larger performances, such as the “Saturday Special Performance” series, it is difficult for Jongmyo Jeryeak to be classified into $F_20$ Performance Work. In addition, as individual records and books related to Jongmyo Jeryeak are difficult to be stored all in the Gugak Archive, they are more often stored in the National Library of Korea, which is a national deposit institution. It is not a common practice for archives to link the data of outside organizations while describing collections in separate archives.

### 8.0 Conclusion

In the Gugak Archive, archival managers are making various attempts to preserve and distribute Korean traditional music. Consequently, the team raised the issues and challenges to be carefully addressed in consideration of the managers’ attempts. Moreover, the characteristics of the gugak and Gugak Archive, as well as the possibility of the web environment and the prerequisites, should be summarized one by one. In addition, the development of online information services provided on the individual websites of the Gugak Archive and the linked data for organizing and sharing the data on the web should be carried out together. In the future, the linked data will demonstrate that it can contribute not only to the Gugak Archive but also to cooperation systems between various institutions that preserve and provide Korean performing arts materials.

### Acknowledgments

We want to express our sincere gratitude and appreciation to all the staff at the Gugak Archive. In particular, we would also like to extend special thanks to Ms. Noh Yeri. This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2016S1A5A2A03927725).

### References


Ziyoung Park – Hansung University, South Korea  
So Young Yoon – Hansung University, South Korea  
Seunghee Son – Hansung University, South Korea  
Yoonwhan Kim – Yonsei University, South Korea

Constructing Semantic Periodical Index Database Focusing on the Visegrad Group’s Transition Process

Abstract:
The purpose of this study is to build a database of articles related to the transition of the Visegrad Group and to use the collected data as a reference for Korean unification. The transitional research team, which studied the Visegrad Group’s histories and political systems, and the database team, comprised of library and information science researchers, collaborated to list articles from 1988 to 2004 and create indexes, summaries, full-text information, and a catalog. The results of this study will be used to design and construct the article index database for multilingual data as well as study the countries’ transition processes.

1.0 Project overview
This research discusses the periodical index database construction of the Visegrád Group, consisting of Central European countries, particularly the Czech Republic, Hungary, Poland, and Slovakia (International Visegrad Fund 2019). The transitional process of these countries during the late 1980s and early 2000s may provide essential insights about Korean reunification.

The project aims to provide relevant information about the Visegrad countries’ transition processes that will serve as a reference for the establishment of a peace regime and reunification of the Korean Peninsula.

The scope of the periodical index database includes the period between 1988 before Mikhail Gorbachev delivered a speech to the United Nations (UN) and 2004 when all the Visegrad countries joined the European Union (EU). This period is divided into three phases:
- Third Phase – Achievement of transition (1999–2004)

This research was conducted by two research teams. First, the transitional research team—comprised of Korean experts on European affairs—studied the Visegrad countries’ modern histories and political systems, discovered and interpreted the relevant articles serving as basic data on the countries’ transitions, built bibliographic information and article summaries, and assigned categories and keywords.

Second, the database team consisted of library and information science majors who can design the data structure, refine raw data, and construct additional structured data to increase the basic data’s usability. The team was responsible for database design, refinement, and quality control and the definition of the metadata types, elements, and values to be built by transition researchers for the overall frame of the database in the early stages of the study. Afterward, all basic data collected yearly were inspected for corrections and authority control. In the interim, the status of the data built by the researchers was monitored to provide additional guidelines or to modify the structure of the initial
database design. In addition, useful external information, such as authority numbers and Wikipedia links, were added to improve the quality and usability of the existing data.

2.0 Initial database design and modification

The major role of the database team was to design the initial database frame. Their second major role was to provide the guidelines for the creation of the metadata of the documents selected by the transition researchers and the quality improvement of basic data from the perspective of information organization. The database was initially designed with the following information:

1. Classificatory information: This includes classification by country based on the magazine’s issuing country, by year based on the year of publication, by source based on the publisher, and by content based on the topic (politics, economy, society, and culture). The place discussed in the content, rather than the magazine’s publisher, will be the location keyword.

2. Descriptive information: The three descriptive elements are elements for the journal, article, and metadata creators and reviewers. Initially, the description included the author’s information; however, it was later removed because of difficulties in acquiring each article author’s name and contact information.

3. Summary information: The summary, written by each country’s assigned researcher, states the content of the full-text article in 100 words. It is written based on facts rather than personal opinions or criticism.

4. Full-text information: The research team gathered all articles in internal storage for metadata reviews and further references.

5. Index information: Index terms were classified into five types by adding general subjects and events to the PLO index, based on the person, location, and organization involved. The importance of index information increases as the project progresses. Transition researchers also assigned the appropriate keywords as they wrote summaries. Then, the keywords, which were written in different languages, were linked with established Korean index terms, and International Standard Name Identifier (ISNI) numbers or Wikipedia link information was added to the proper names of people, locations, and organizations.

3.0 Document selection and acquisition of full-text articles

In this process, history and political science majors, who are capable of understanding local languages and analyzing each country’s historical events, first selected, analyzed, and described the articles. They created the basic descriptive catalog, wrote summaries in Korean, and assigned keywords and categories for each article. For article selection, they examined 10 kinds of daily and weekly journals issued in the transition countries— the Czech Republic and Slovakia (Czechoslovakia), Poland, and Hungary—and the countries that influenced them, Germany and the US.

The researchers checked and selected the full text of all articles, while considering each country’s research period and equity, through local visits and web services. It was difficult to select only a few hundred articles per journal among the multitude of articles within a limited time. The final number of articles collected by country is shown in Figure 1.
4.0 Basic metadata creation and refinement

At the beginning of the study, the database team outlined the initial design of metadata elements and values and provided the transition team with guidelines to build their basic metadata—the bibliographic information, summaries, and keyword lists of the articles in Korean. After creating the metadata and descriptive catalog, the researchers wrote each article's Korean summary, which is 100 words long, that can help Korean researchers in searching and understanding the articles in other languages. The researchers also assign keywords to each article, which are classified by the person, location, organization, general subject, and event involved. In addition, each article is also categorized into politics, economics, society, and culture.

Once the transition researchers complete the article selection and the primary metadata, the database team checks the metadata from the full-text image files and ensures that it meets the criteria stated in the guidelines. Then, the database team corrects any identified errors and asks transition researchers for feedback when necessary. Besides spelling errors, the most common types of errors are the following:

- Titles and subtitles selection: The main title, usually in the largest font size, is the main title of the article. On the other hand, the subtitle, consisting of phrases that expound on the title, is usually presented in a larger font on the top or bottom of the main title. However, there were several instances when they were mistakenly switched.
- Incomplete full-text article: Parts of the original text was not obtained when newspaper or magazine articles begin on the first page but are continued on another page.
- Other basic errors: Incorrect capitalization in titles, missing volume, and issue information for articles, and incorrectly copied original link information available on the web.

The following are examples of summarized metadata for each article:

- Descriptive elements for the unit of journals

Figure 1. Number of articles collected by country
Some of the negotiation themes to join the European Union (EU) are not in the EU regulations; therefore, some topics would not be a problem unless they endanger the EU. One of the more difficult areas is the Schengen regulatory system, particularly for topics such as changing the monitoring and control of borders that violate EU standards. (Omitted below) [Translated from the Korean summary]

- Article category: “경제” (Economy) [Korean]
- Article keywords (Korean established/original keywords):
  - [organization] 유럽연합 (European Union)
  - [event] 유럽연합 가입협상(EU csatlakozási tárgyalás)
  - [event] 유로화 도입 (Euro bevezetése)
  - [general subject] 셰겔규정 (Schengeni előírások)
  - [general subject] 국경통제 (határellenőrzés)

© Ergon - ein Verlag in der Nomos Verlagsgesellschaft
5.0 Data enhancement by constructing index term clusters

In addition to correcting data errors, the database team also examined article classification and indexing information, which consists of author keywords and controlled index terms. The index terms are in different languages, such as Czech, Polish, Hungarian, German, and English, and the team has made a cluster by gathering words with similar meanings and then adding the preferred Korean term.

Semantic grouping of concepts is essential for semantic search capabilities (King and Reinold 2008, 27). If individual terms are divided into facets, the index term list can be called “faceted” (Chan and O’Neill 2010, 13). Given the articles’ data written in Czech, Polish, Hungarian, German, and English, the database team constructed a faceted multilingual index term cluster by grouping multilingual index terms for each preferred Korean index term. Moreover, the team assigned a specific identifier to each index term cluster by concept and linked them to other entities. The following points were considered in building a multilingual index term cluster:

- Similar index terms constructed by language are grouped according to the established Korean index term and examined by experts of each language.
- In the case of proper names, such as institutions or organizations, even if the Korean translation was the same, the terms are considered different.
- After drafting a multilingual cluster, another researcher in the team cross-checked the consistency of index clusters within the same language.
6.0 Data enhancement by adding controlled link information

The database team also added external link information to the index term clusters to strengthen semantics because Koreans might be unfamiliar with the persons, locations, organizations, events related to the Visegrad Group’s transition. The database includes external link information created by adding Uniform Resource Identifiers (URIs) of related bibliographic entities from reliable sources, such as ISNI, the National Library of Korea, and Wikipedia, as well as index terms clusters that combine the keywords assigned by researchers in each magazine.

The team designated Korean Wikipedia links and the link information provided by Linked Open Data from the National Library of Korea as external links for all index terms. Furthermore, the team added ISNI to the Person and Organization index terms, and URI information from the GeoNames site to the Location index terms. External link information is also useful, especially for the semantic web environment and information sharing (Lee, Park, and Lee 2017; Park 2012).

7.0 Conclusion

In this study, articles related to the system transition of the Visegrad Group were selected from representative weekly and daily magazines to establish the article index database. In the process, the team systematically designed the catalog, classification/index, summary, and fulltex information. Through repeated data reviews, the team corrected basic errors and built index clusters so that articles in multiple languages could be semantically linked. Key figures and events related to the transition also provided additional information in conjunction with ISNI or Wikipedia pages. Index term and descriptive information about each article are crucial in the periodical index database. Index terms were found to improve articles’ findability and help users understand the articles and locate relevant information. Moreover, with the constructed faceted index terms,
articles from a certain period can be linked to a related person, organization, location, and event. Well-arranged sets of multilingual index terms also enabled the use of a reliable multilingual dictionary. Furthermore, the links to ISNI, Wikipedia, GeoNames, and Authority Data from the National Library of Korea helped provide additional person, event, and bibliographic information. The periodical index database constructed through this research can be used as a basis for Korean unification studies and contribute to further research about unification and transition. In addition, the data and the skills acquired during the construction process will also be used as a reference for creating and using an article index for multilingual data in the future.

Acknowledgments
This research, financially supported by Hansung University, is based on the Fundamental Research Program “Compilation of the Documents regarding the Democratic Transformation of Eastern Europe: 1989–2004: With a Special Focus on the Visegrad Four’s Periodical,” supported by National Research Foundation. We want to express our sincere gratitude and appreciation to all the researchers who participated in the project.

References
Facilitating University–Industry Interaction by Visually Showcasing Researcher Profiles Via Metadata

Abstract:
University–industry interaction (UII) has grown significantly over the past two decades. Yet, contacting universities to establish collaboration continues to be perceived as challenging by the industry. This paper studies the challenges associated with UII for analysing how metadata, taxonomies and visualizations in research information management systems (RIMS) may support UII for large enterprises. The first step was a systematic literature review conducted to understand the challenges associated with UII. The second step was interviews with R&D managers from three large enterprises to extend our knowledge about this industry group. The results showed that the primary channel for establishing UII was through personal connections. The informants saw the RIMS as an important tool for browsing research literature and exploring research groups to gain insights into research topics and individual researchers. In terms of metadata, the findings showed that multi-disciplinarity and high-level granularity were important aspects. Furthermore, the visualisation of relationships and description of international collaborations was perceived a useful indicator of researchers’ overall quality and impact. Similarly, metadata describing job titles, departments, and citations was central for judging the credibility of experts. An interesting finding was that enterprise managers found it difficult to develop personal relationships with relevant academic experts. Future studies may benefit from interviews with HR management professionals exploring how to support recruitment by using metadata and taxonomies.

1.0 Introduction
This study investigates the roles of taxonomy and metadata in showcasing researchers’ qualifications by using research information management systems (RIMS) as a part of university–industry interaction (UII). In addition, we analyse how RIMS can be developed to visually showcase researcher expertise and explore how researcher profile data can be supported with taxonomies and metadata to better facilitate UII. Generally, UII is defined as the “interactions between all parts of higher-educational systems and the industrializing economy” (Ankrah et al. 2012, 50). UII has greatly increased over past two decades owing to an increase in federal funding for research and development (Azeroual, Saake, and Wastl 2018). RIMS are considered key tools for UII, and to design and manage RIMS, it important to not only update descriptions of experts’ knowledge, skills, and impacts on an ongoing basis but also to provide high-quality data to ensure that RIMS can be used to support various user groups (Ebert et al. 2015). Data visualization is ‘speeding up’ the cognitive processes of filtering information; therefore, it plays an important role in the sciences as a method for generating insights (Fekete et al. 2012).

A growing number of studies have addressed the challenges associated with UII (Penfield et al. 2014). Recent scientific studies have outlined the diverse challenges facing UII, for example, finding partners for collaboration and contacting universities (Freitas, Geuna, and Rossi 2013). Another core problem is representing the specific knowledge and expertise in researcher profiles (Ehrlich 2003). It has been suggested that the needs and perspectives of large enterprises deserve special attention because such enterprises play a critical role in the world economy as innovators (Ebert et al. 2015). The gap
between universities and the industry and the lack of knowledge in academia about large enterprises led us to formulate the following research questions:

RQ1: What challenges do large enterprises face in terms of UII?
RQ2: How can we design metadata schemes and taxonomies so that they can provide researcher profile data that fulfil the needs of large enterprises?
RQ3: How can we graphically present researchers’ profile data to enterprises to facilitate UII?

2.0 Research methods

This study consists of two sub-studies: a literature review of the challenges associated with UII as they appear in previous studies and an interview study that extends and specifies the challenges faced by large enterprises and their viewpoints on metadata, taxonomy and visualization. RQ 1 is answered using the data obtained in the two studies. RQs 2 and 3 are answered using the data obtained in the two studies and the theoretical literature on information architecture (Morville, Rosenfeld, and Arango 2015) and visualization (Steele and Iliinsky 2010).

3.0 Literature review

This sub-study was carried out as a systematic literature review (LR), and the methodological process employed for this LR was inspired by Ridley (2012). The LR was performed with the following aims: 1) provide an overview of UII characteristics; 2) understand the motives of UII; 3) discover the challenges related to UII from the industry’s perspective; 4) learn about the channels used to establish UII; and 5) understand how information should be presented to the industry. Titles that included terms such as UII, university–industry collaboration, and challenges or barriers to UII or university–industry collaboration were selected for inclusion in the literature review. Publications that included empirical data were prioritized because personal experiences are important in phenomenological studies (Lester 1999). In total, 27 publications were selected. A few of these publications were eliminated because they studied UII primarily from a university perspective or because they were considered unreliable sources or were not peer-reviewed. Only 20 publications were reviewed in detail, and among them, only 8 were included in the actual analysis. During the reading process, the texts were colour-coded to help address the synthesis among the selected articles (Ridley 2012). The reviewed literature covered UII from the perspectives of large and small and medium-sized enterprises.

4.0 Interview study

Three large-sized engineering enterprises based in Northern Jutland were selected for the case study. These three enterprises fit the definition of large enterprises, that is, enterprises with more than 250 employees (Løkkegaard 2018). The interview study consisted of three interviews with research managers, one from each of the enterprises. The field of engineering was selected because it is one of the leading areas for UII (Murashova and Loginova 2017). The target sample was identified using a combination of convenience and purposive sampling methods (Bryman 2016). The main selection criteria were as follows: Informants should have professional functions related to universities
or research activities, and UII should be relevant to the informants’ companies. Informants 1 and 2 had dual positions: they worked as research managers with their respective companies and as part-time industry professors with Aalborg University. Informant 2 was with Aalborg University as a supervisor for master’s students. For details, see Figure 1.

<table>
<thead>
<tr>
<th>Informant 1</th>
<th>Industry name and type</th>
<th>Educational background</th>
<th>Work title/functionalities</th>
<th>Relationship with University</th>
<th>Gender</th>
<th>Enterprise size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informant 1</td>
<td>Mechanical engineering, water system solution development</td>
<td>M.Sc., Control Engineering and Automation, Ph.D. in Control Engineering</td>
<td>Chief Engineer/Chief Specialist, working with control and supervision systems both at Grundfos and Aalborg University</td>
<td>Industry Professor, part-time Professor at Department of Electronic Systems, The Technical Faculty of IT &amp; Design, Automation &amp; Control</td>
<td>Male</td>
<td>&lt;19,000 employees globally</td>
</tr>
<tr>
<td>Informant 2</td>
<td>Robotic optimization development for industry</td>
<td>M.Sc., Mathematics &amp; Computer Science</td>
<td>Product Manager: business development, concept development and product management</td>
<td>No university-related position</td>
<td>Male</td>
<td>&lt;370 employees in Denmark, Sweden and Norway</td>
</tr>
<tr>
<td>Informant 3</td>
<td>Manufacturer of electronics and audio products, television sets and telephones</td>
<td>M.Sc. and Ph.D in Acoustics</td>
<td>Director Research, responsible for research activities and managing research groups</td>
<td>Industry Professor, part-time professor at The Technical Faculty of IT and Design, Electronic Systems, Signals &amp; Information processing</td>
<td>Male</td>
<td>&lt;1,028 employees</td>
</tr>
</tbody>
</table>

Figure 1. Information about selected sample and industry.

The RIMS and the related metadata scheme and taxonomy used by Aalborg University were presented to the participants as an exemplary RIMS system during the interviews. This RIMS was selected because it supports the visualization of academic expertise, networking and collaboration. The taxonomy and related metadata were discussed and evaluated together with the visual presentation of researcher profile data. Several graphics were discussed during the interviews, for instance, visualizations generated using a fingerprint algorithm that captured subject terms from uploaded abstracts and network visualizations displaying the relationships between researchers and academic departments, as shown in Figures 2 and 3 (Elsevier 2016).
The interviews were exploratory and open ended, and the informants were asked questions about the following themes: 1) understanding the contexts of the informants and their perceptions of UII and RIMS; 2) evaluating the categories and metadata in the RIMS; 3) evaluating graphic visualizations in the RIMS; and 4) gathering insights as to whether the RIMS can facilitate UII. The interviews were started by asking the informants to sign a consent form, and they were conducted in natural settings with the informants sitting in their offices. The interviews with informants 1 and 2 were conducted in person at their offices, while the interview with informant 3 was conducted online, wherein the informant sat in his office and the researchers in the university. The researchers shared their screen with the informant during the interview to allow the informant to explore the RIMS. The interviews were recorded and transcribed, leading to more than 160 min of interview video in total.

A thematic meaning condensation process was used to analyse the transcriptions, and commonalities, relationships and differences across the data were identified (Gibson and Brown 2009). The exploratory process of meaning condensation was used because it provided the researchers with reflective and detailed steps to conduct data analysis (Malterud 2012). A ‘meaning unit’ is defined as a fragment of text containing some information relevant to the research question (Malterud 2012). Inductive reasoning was adopted to conduct the analysis. NVivo computer software was used to perform the analysis (NVivo 2020).
5.0 UII characteristics and challenges

The LR findings confirmed a lack of studies focused on investigating UII challenges, for example, challenges related to the identification of experts by using the RIMS. The majority of the studies focused on industry perceptions of the university (Vick and Robertson 2017). The review showed that the primary motives for UII were ‘rising costs’ and ‘societal constraints’. On a personal level, academics were self-oriented to engage in UII for creating career opportunities within their organizations (Vick and Robertson 2017). Surveys conducted in 9 academic departments across 115 universities suggested that researchers are careful in establishing UII because it restricts their academic freedom. Moreover, the surveys indicated an underlying tension for research funding that weighed against the need for academic freedom (Ankrah et al. 2012). The need to find and hire talented students was a motivating factor for the industry to establish UII (Ankrah and AL-Tabbaa 2015). Additionally, the desire to attain ‘competitive advantages’, ‘stability’ and ‘legitimacy’ were further motives for establishing UII (Ankrah et al. 2012).

Large enterprises have a large absorptive capacity for using scientific knowledge and better capabilities in terms of searching and identifying knowledge providers (Ankrah et al. 2012). According to a study, the main challenges associated with establishing effective UII were organizational differences between universities and the industry (different aims, levels of formality, risk perceptions and values) (Collier, Gray, and Ahn 2011). Similarly, enterprises were found to perceive universities as ‘a different working environment’ (Ankrah and AL-Tabbaa 2015). Quality issues were identified as a challenge, and academics were perceived to be too theoretical and not very practical, whereas the industry’s focus is considerably more practical with a centred interest on critical issues (Ankrah et al. 2012). Moreover, companies find it difficult to approach and make contact with universities (Ankrah et al. 2012). Enterprise managers find it difficult to ‘identify skills, their firms needed and then to develop personal relationship with academic experts’ (Ankrah and AL-Tabbaa 2015).
Our analysis of the interview study revealed similarities between the findings of the LR and those of the interviews, such as enterprise size and its effect on absorptive capacity (Freitas, Geuna, and Rossi 2012). However, the interview findings provided more insights and unexpected findings. Informant 1 described how they perceive their company as a ‘look-alike-university’ and are familiar with scientific knowledge. Moreover, the interview findings suggested that within enterprise 3, in-house research groups are formed, and these groups tend to solve problems internally. This finding was unexpected because according to the LR, large enterprises tend to use collaborative research programs rather than in-house research programs to reduce costs (Ankrah and AL-Tabbaa 2015).

Another finding was that for large enterprises, it is important ‘to collaborate with people that they know’ (Informant 1). Similarities were found with regard to the LR findings because in some cases, it is not the expertise or the qualities of an expert that matter but rather the ‘personal traits of the academic expert’ (Collier, Gray, and Ahn 2011). Another similarity between the findings of the LR and the interview study was that enterprises use ‘personal contacts’ as a channel for establishing UII (Ankrah and AL-Tabbaa 2015). The LR suggested that the challenges faced in establishing effective UII were the ‘difficulty to make contact with the university’ and identification of the right partners’ (Freitas, Geuna, and Rossi 2012). However, a surprising discovery that was disproven by the interview study was that finding an expert for collaboration and contacting a university were not perceived as problems by the informants. One explanation may be that the three informants maintained close collaborations with Aalborg University (AAU) through internship programs, as well as the close connections between the university and surrounding enterprises owing to the AAU tradition of problem-based learning, which encourages students to collaborate with enterprises in their project work to work on real-life cases (Aalborg University 2020).

6.0 Metadata, taxonomy and graphics in RIMS

The LR study revealed that the information presented to an enterprise should be ‘easy to use’, ‘practical’, ‘visually attractive’, ‘short and specific’ and ‘quickly decoded’ (Løkkegaard 2018). Visual aspects such as an expert ‘profile picture’ are important, and the RIMS should be supported with taxonomy to ensure that it represents user needs (Ehrlich 2003). Moreover, it was suggested in a study that granularity provides one with the ability to rank experts by using narrower criteria and showcasing multiple relationships among experts, co-authorship, citation links and project groups (Yimam-Seid and Kobsa 2003). Enhancing the RIMS with ontology-based presentations of expertise, such as subset-superset relationships and multidisciplinary and interdisciplinary expertise, considerably increases the probability of selecting the right expert (Yimam-Seid and Kobsa 2003). Similar findings from the LR were that ‘expertise’ and ‘subject categories’ should be supported by a highly granular taxonomy and that metadata should contain ‘credentials’, ‘accessibility’ and ‘demographics’ of the profiled researchers (Ehrlich 2003). Moreover, it must be possible for one to assess an expert’s credibility by reviewing their ‘published papers and awards’, ‘grants and patents’ and ‘professional affiliations’ (Ehrlich 2003). During the interviews, it was found that when looking for academic experts, the informants started their search in the ‘publications’ category.
Informants 1 and 3 explained that they looked for topically relevant publications to identify universities and research groups that were publishing within the field of interest and could, thus, be considered relevant for their enterprise. They searched by ‘concepts’ and ‘topics’. The specific topics that represented their field and domain vocabulary were important and relevant to the informants. This was reportedly the first step towards finding relevant experts.

Descriptive metadata about publications, such as title, author and abstract, were found to be important. The abstract helped the informants judge whether a subject was interesting and whether a publication was theoretical or application-oriented. The second step was to quickly identify the ‘leaders’ of the publications or members of the ‘research group’ because these data provided information about the overall quality of the research/researchers. The three informants found that citation indexes presenting productivity and metrics provided useful descriptions that helped them identify the level of activity of a researcher and whether a researcher was a professor or a PhD student. Moreover, such indexes ‘help identify the key supervisors or leaders of a research group’ (Informant 1). Similarly, the h-index helped ‘identify the key scientific persons in this topic area’ (Informant 3). Commonly, PhD students would have a lower index. This means that citation indexes and the h-index were considered useful sources of metadata that helped the informants assess an individual’s level of expertise. The informants stressed that metadata about research activity on specific topics across departments and international collaborations should be included in the RIMS. The findings of Løkkegaard (2018) supported the interview findings that subject information about researchers’ scientific knowledge is important for enterprises. She added that scientific knowledge should be presented such that it is clear how the knowledge can be used and applied in practice. Informant 2 stated that graphical expertise exposure was important and useful because ‘it helps the memory’ and assists with ‘interpretation of information’. Profile pictures were considered important because they ‘give an idea of what a person is’. According to Informant 1, visual presentations ‘catch attention’, thus confirming that visual exposure is important and supports cognition. Visual information is considerably easier to perceive than textual information (Shneiderman 1996). Likewise, the availability of researcher profile pictures in expert-finder systems was found to be important (Yimam-Seid and Kobsa 2003a).

According to the LR, an unusual finding of both the LR and the interviews was that enterprise managers found it difficult to ‘identify skills their firms needed and then to develop personal relationship with academic experts’ (Ankrah and AL-Tabbaa 2015). Informant 1 explained how network visualization could be useful from the perspective of staff management because it would help one plan which are the ‘places where we want to work, which topics to work on and who we want collaborate with’.

To sum up, in most cases, the interview participants found useful the metadata and taxonomy used in the Aalborg University Pure Portal. The ‘publications’ category was relevant from the viewpoint of searching for experts, and the ‘subject’ category allowed the informants to search the Pure Portal for information by automatically generating ‘concepts’ and ‘topics’ as categories. Metadata describing the ‘title’, ‘department’ and ‘individual and research unit collaborations’ were deemed useful. Furthermore, the ‘citation index’ and ‘h-index’ were found to provide useful information about the credibility of an expert. However, some taxonomy terms and metadata types were missing. The
interview participants missed metadata describing the ‘leader’ of a research group, which would be useful for identifying the leader of a publication. Metadata related to and describing ‘international collaborators’ and ‘departmental collaborators’ were considered useful information as well. Likewise, it was found that metadata on external individuals who have collaborated with an expert would be useful for network visualization. The findings further suggest that for graphic visualizations to facilitate UII, they must be informative and communicate the information that is relevant to a user. When designing a visualization, it is important to prioritize information over a superfluous design that can confuse the user. To ensure that the visuals remain relevant, the designers should achieve a balance between novelty and efficiency; in other words, redundancy should be minimised, so that the intended meaning is not lost in pursuit of a highly unique design. Likewise, visualizations should present the use context and information in an effective manner to support UII. The study found that graphic visualizations that can be explored (provide relationships to different datasets) are more relevant than those that do not facilitate exploration. Similarly, visualizations that employ graphic elements for enhancing important information (by using colours or bold characters) are perceived as useful and more efficient. Minimizing visual clusters by reducing redundant graphical elements, such as lines or numbers, may help make a visualization more aesthetically attractive.

For graphic visualizations to be informative, they must be supported by useful metadata that expresses a clear, unambiguous meaning and showcase metrics that can answer users’ questions. To support visualizations, metadata terms must be specific, clearly describe the intended message and provide metacommunication to support the context of use. The taxonomy vocabulary should avoid ambiguous terms and should maintain specificity and domain-orientation. Simultaneously, the taxonomy must provide multidisciplinarity, in addition to showing and relating perspectives and vocabulary from a set of relevant domains.

8.0 Conclusion
This study explored metadata, taxonomy categories and graphic visualizations as means to support descriptions of researcher expertise in RIMS. This showcase aims to solve the challenges associated with UII, and consequently, improve the interaction between universities and the industry. Moreover, the study explored UII from the perspective of large enterprises. The LR findings suggested that when searching for academic knowledge and researchers, industry professionals prefer graphic visuals instead of textual information because visuals catch their attention and are easy to understand. Moreover, enterprises find it challenging to establish collaborations with universities. Therefore, the study aimed to understand how to better present researcher expertise with metadata and a granular taxonomy to resolve the challenges associated of finding the right experts for establishing UII. Taxonomy and metadata provide context, consistency and information regarding visualizations and help enterprises to determine the level of expertise and cross-departmental collaboration of an individual. A prototype RIMS was used as a typical identification case. The essential finding of this study is that establishing UII was not perceived as a problem by the informants, which contradicts the LR findings. This contradiction was ascribed to the informants’ close personal connections with the university. Personal contacts were found to be the most efficient channels for
interacting with researchers, as opposed to RIMS. Moreover, visualizations were found to be important and useful for enterprises because they provide a ‘quick interpretation of information’. Metadata should support the descriptions of visuals and provide meta-communication regarding the context of use and how the visuals were generated. Multidisciplinary descriptions are important for the industry, and therefore, taxonomy should include high-level granularity and domain-specific terminology to support information.

In sum, the findings provided an understanding of UII from the perspective of large enterprises in relation to graphic exposure, taxonomies and metadata in RIMS. However, it is important to stress that improving the taxonomy and metadata would not necessarily change the ways in which industry actors establish collaborations with academic experts, because collaborating with people who are personal connections is still the preferred route for establishing interactions. Future studies should explore industry professionals in management positions to explore how metadata can support recruitment and business development.

References


Morville, Peter, Louis Rosenfeld, and Jorge Arango. 2015. Information Architecture: For the Web and Beyond. Sebastopol, CA: O’Reilly Media, Inc.


© Ergon - ein Verlag in der Nomos Verlagsgesellschaft
Günter Reiner – Department of Law, Helmut Schmidt University, Germany
Philipp Adämmer – Department of Mathematics and Statistics, Helmut Schmidt University, Germany

Similarities Between Human Structured Subject Indexing and Probabilistic Topic Models

Abstract:
This paper adds statistical findings from natural language processing to an ongoing interdisciplinary research project between lawyers and information scientists. The project proposes an indexing scheme that follows a content grid of six predefined categories, also called facets in a broad sense. These facets try to capture the essential structure of legal information and are based on Ranganathan’s fundamental facets as well as on the famous Roman lawyer’s Gaius tripartite division of persons, things and actions. The prototype database, built as part of the project, consists of nearly 2,500 cases which have been manually indexed. The present study uses the prototype database to investigate whether similarities exist between human subject indexing and automatically clustered terms. We first examine the similarities between the indexing terms and terms generated by a term frequency-inverse document frequency approach. Then we investigate the similarities between the indexing terms and word clusters generated by unsupervised probabilistic topic models, namely latent Dirichlet allocation (LDA) and the correlated topic model (CTM). On average, the similarities of the manually indexed terms with the topic terms are not high but statistically significant, and for some cases we find strong similarities in which the clustered terms of the topic models match well the humanly indexed terms. Correlations are slightly higher when using the CTM instead of LDA. Our results indicate that topic modelling could be beneficial at least for semi-automatic indexing. Disentangling further which facets contain those terms that predominantly cause similarities with automatically clustered terms could further enhance the support for human structured subject indexing.

1.0 Introduction
Indexing is one of the oldest tools to foster the retrieval of written information. Even in the age of electronic full-text searches, it has proven its worth (Gross et al. 2015). The present project ties in with an ongoing project between lawyers and information scientists, which is funded by the Social Sciences and Humanities Research Council in Canada and in which one of the authors is involved (Cumyn et al. 2019). The Canadian project proposes an indexing scheme that follows a content grid of six predefined categories, also called facets in the broadest sense. These facets (Person, Action, Thing, Context, Legal category and Sanction) attempt to capture the essential structure, so to speak the “grammar” (Reiner et al. 2019, 352-353), of legal information in two ways: first in terms of the fundamental division between the factual elements of a case (the first four facets) and the legal consequences (the last two facets). Second, in terms of the way factual information is analysed, which is based on Ranganathan’s universal facets (Personality, Matter, Energy, Space, Time) and the famous Roman jurist Gaius’ famous tripartite distinction of persons, things and actions (Cumyn et al. 2018; 2019; Reiner et al. 2019).

A test database (called “Gaius”), which is an excerpt from the offering of the commercial Quebec database operator SOQUIIJ (Société québécoise d’information juridique), has been created as part of the Canadian project. It contains 2,500 court decisions from Quebec (mostly in French) which are divided into five sub-databases according to those fields of law that SOQUIIJ had assigned to them (administrative law, labour law,
contract law etc.). These decisions were manually re-indexed using a faceted scheme on the basis of a controlled vocabulary (thesaurus) being developed gradually and kept as lean as possible.

Human indexing is expensive and time consuming, which is why we attempt to find methods for semi- or even fully automatic indexing. It is not linked to the Canadian project in terms of its research goal, yet it is based on the test database described above, that is an interesting source of information, since it allows to compare human (manual) indexing with automated indexing. Legal databases that are systematically indexed according to content criteria are rare. In addition, the Gaius database has the advantage that the indexing is structured, which enables specific statistical analyses revealing indications of human indexing patterns. Probabilistic methods of computerized text analysis are more similar to the human understanding of (legal) texts than one might think; it is well-known that the natural acquisition and processing of language is based not on the application of rigid rules but on experience, which can be simulated using an inductive process based on the estimation of probabilities (Chater and Manning 2006, 340). Our project aims at investigating whether probabilistic topic models, such as latent Dirichlet allocation of Blei et al. (2003), create and assign word clusters to legal documents (court decisions) in a manner that is similar to human facet indexing. According to the literature, there has been experience with topic modeling in the field of legal information for several years (George et al. 2014; Livermore et al. 2017), but so far they remain isolated and do not serve the purpose of indexing.

We are approaching the purpose with the following sub-questions:

1. Are the similarities between human indexing and automatically generated keywords generally higher when using the words of the topic models instead of the words generated using a frequency–inverse document frequency approach?

2. Is there a statistically significant relationship between human indexing and keywords generated by topic modeling? If so, topic models can be useful for (semi-)automatic indexing.

3. Do facet indexing and topic keywords created by topic models correlate more strongly than unstructured indexing (e.g., SOQUIJ indexing) and topic modeling keywords? If so, this could be an indication that faceted indexing - beyond its supposed advantages for queries - offers advantages in (semi) automation.

4. Are there some of the six facets which, by the terms assigned hereto, predominantly cause similarities with automatically clustered terms? If so, this insight could help to enhance human structured subject indexing (conceptually and in individual cases) up to semi-automatic indexing.

2.0 Empirical analysis

To test the hypotheses, we have automatically generated word lists using the term frequency–inverse document frequency approach (2.1) and the topic model approach (2.2). We compared the results with the manual indexing of the Gaius database using cosine similarities (2.3).
2.1 Term frequency–inverse document frequency

The term frequency inverse document frequency (tf-idf) aims to measure the relevance of a word within a certain document. The term frequency (tf) simply counts how often a word (hereafter \( w \)) occurs in a document (hereafter \( d \)). Yet the informational content of terms that occur frequently in many documents (e.g., the, and, for, . . . , etc.) is mostly low. It is rather those terms that appear frequently in a small number of documents but rarely in the other ones that tend to be informative (Huang, 2008, p. 51). The tf-idf accounts for this aspect whose formula can be written as follows:

\[
tf-idf(w, d) = tf(w, d) \times \log \left( \frac{N}{df(w)} \right),
\]

where \( tf(w, d) \) is the term frequency of \( w \) in \( d \) and \( df(w) \) denotes the number of documents in which \( w \) occurs. \( N \) equals the number of documents in a corpus and \( \log() \) is the logarithm with base 10. Terms that occur frequently in some documents but only rarely in the overall corpus yield high tf-idf values. We computed tf-idf values for each word in each legal document of the sub-databases Admin (administrative law), Contrats (contract law) and Travail (labor law). The legal documents are grouped according to their sub-database. We used separately the 10, 20, . . . , 50 words of each document with the highest tf-idf values for comparison with the corresponding index terms from the Gaius database.

2.2 Probabilistic topic models

Probabilistic topic models (TM) are algorithms for the analysis of large document collections. In contrast to the tf-idf approach, TM can assign terms to documents that are not included in the document itself (cross-referencing). In addition, TM assume that documents are written by a stochastic process in which all documents share \( K \) common topics. A topic is a discrete probability distribution over words. All topics contain the same words, namely the totality of all words in the database, but the probabilities given to each word differ. For example, a topic about damages would give high probabilities to words such as negligence and causation, while a topic about labor contracts would put high probabilities on words such as employee and dismissal. Each document is then assumed to be a mixture of those corpus wide topics. The topic mixture for each document is given by the so-called topic proportion.

The most popular and most cited TM is latent Dirichlet allocation (LDA) by Blei et al. (2003). The model owes its name to the fact that the topics and the topic proportions are assumed to be drawn from a Dirichlet distribution. Each element of a randomly drawn Dirichlet vector is between zero and one. In addition, the elements of a Dirichlet vector sum up to one, thereby complying with the requirements of probabilities. However, one drawback is that the Dirichlet distribution cannot account for correlations. For example, if a legal document writes about contracts it is more likely that it also deals with frustration than if it deals with administrative law. Therefore, we also used the

---

1 The base can certainly be changed, but we decided to stay with the default settings given and justified in the R-package quanteda by Benoit et al. (2018).
correlated topic model (CTM) by Blei and Lafferty (2007). As is common in natural language processing, we removed a whole bunch of stop words (e.g. aux, notre, nous, que), hyphens, apostrophes, numbers, etc., before estimating (computing) the models. We also removed words that have been classified as irrelevant by ourselves (e.g., demande). To avoid the particular difficulties of a multilingual database, we have also excluded automatically the few English language decisions. Finally, we replaced certain words in the documents according to the synonym list of the Gaius thesaurus. Although being unsupervised learning algorithms, LDA and the CTM require the selection of the number of topics $K$. Both TM were estimated with $K = 10, 20, \ldots, 50$. For each $K$, we assigned successively the 1, 2, 3 most probable topic(s) to each document (indicated by the topic proportions). We then chose successively the 5, 10, 15 most probable words of each topic. With a fixed number of topics $K$, we thus had 9 (3x3) combinations of terms to compare with the human indexing for each legal document.

2.3 Measuring similarities with human indexing

To measure how similar the words from the automated approach are to the humanly indexed terms of the Gaius database, we used the concept of cosine similarities (see, e.g., Huang 2008). In the first step, we converted the total Gaius indexing and the automated word lists into a single so-called document-term matrix (dtm), where each row ($i$) corresponds to a classification of one legal document (Gaius or automated) and each column ($j$) denotes a unique word. Each cell entry thus indicates how often the word $j$ occurred in the classification document $i$. A classification document can thus be represented as a vector in high dimensional space.

The idea of the cosine similarity is to measure the angle between two vectors. In our case, it measures how close the classification vectors from the Gaius database and each of the corresponding automated approaches are. On the one hand, if two vectors are identical, the angle between them is zero. The cosine of zero equals one; this value therefore represents the highest similarity. On the other hand, if the two vectors are orthogonal to each other (no words intersect), they have an angle of 90°, whose cosine value equals zero. We thus have a measure bounded between zero and one that indicates how close the Gaius indexation and our automated word lists are.

The cosine similarity between two vectors can be computed as:

$$
\cos(\theta) = \frac{V_G \cdot V_M}{\|V_G\| \|V_M\|} = \frac{\sum_{w=1}^{Z} V_{G,w} \cdot V_{M,w}}{\sqrt{\sum_{w=1}^{Z} V_{G,w}^2} \cdot \sqrt{\sum_{w=1}^{Z} V_{M,w}^2}},
$$

(2)

It does not make sense to try to eliminate all irrelevant words in advance. This is not only time-consuming but above all, there are many terms whose (missing) relevance to the document content depends on the context.

Choosing the optimal number of topics depends on the purpose of the analysis. For example, predicting unseen documents is a different task than trying to find the optimal number of semantically meaningful topics for a fixed set of documents. Several metrics have been proposed in the literature to find the optimal number of topics $K$ (see, e.g., Roberts et al. 2019). We used the R-package topicmodels by Grün and Hornik (2011) to estimate LDA and the stm package by Roberts et al. (2019) to estimate the CTM.
where $G$ and $M$ denote the vector of word counts for the Gaius database ($G$) and the automated (tf-idf or TM) word lists ($M$). The number of unique words is given by $Z$. The nominator computes the dot product of two vectors and the denominator is the product of the vectors’ Euclidean norms. Since the Gaius database uses a lot of compounded words for indexation (i.e., *permis d’alcool*), we have split them into separate strings to make the terms comparable. We also broke down the words to their stem, such that, for instance, words in plural are given in singular form.

### 3.0 Empirical Results

#### 3.1 Quantitative Results

Table 1 shows an example of the five most probable words of the seven most prevalent topics in the sub-database *Contrats*. The topics were estimated with the CTM and the total number of topics was 20. As the topic model randomly assigns topic numbers, we renumbered the topics for illustration purposes from 1 to 7.

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
<th>Topic 6</th>
<th>Topic 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>somme</td>
<td>malfaçon</td>
<td>travaux</td>
<td>véhicule</td>
<td>somme</td>
<td>vente</td>
<td>preuve</td>
</tr>
<tr>
<td>preuve</td>
<td>acheteur</td>
<td>preuve</td>
<td>vente</td>
<td>contrat</td>
<td>immeuble</td>
<td>contrat</td>
</tr>
<tr>
<td>services</td>
<td>vente</td>
<td>contrat</td>
<td>prix</td>
<td>prix</td>
<td>contrat</td>
<td>somme</td>
</tr>
<tr>
<td>contrat</td>
<td>eau</td>
<td>dommages</td>
<td>garantie</td>
<td>travaux</td>
<td>promesse</td>
<td>être</td>
</tr>
<tr>
<td>payer</td>
<td>être</td>
<td>somme</td>
<td>bien</td>
<td>argent</td>
<td>représentant</td>
<td>droit</td>
</tr>
</tbody>
</table>

The terms of the TM were used to compute the cosine similarities in the manner as described above. Figures 1 and 2 show estimated kernel densities for the cosine similarities between indexed terms of the Gaius database and words created by both text mining approaches (tf-idf and TM). The overall area of each estimated density sums up to one. The further the density area is shifted to the right, the more often occur higher cosine similarities. We have chosen the number of words per topic and the number of topics assigned to each document so that (i) the standard deviation of all cosine similarities is the lowest (shown in Figure 1) and (ii) the average of the cosine similarities is the highest (Figure 2). With regard to the optimal number of topics, this resulted in a uniform value of three per document.

The figures show that the cosine similarities between words of the TM models and the Gaius indexing are higher than those between the words given by the tf-idf and the Gaius indexing. This visual impression is statistically confirmed by t-tests regarding the mean values. Additional t-tests on the differences of cosines similarities between LDA and the CTM indicate that the CTM approach yields higher cosine similarities or, put differently, CTM words are, on average, more similar with the human indexed database.

---

4 The table shows the five most probable words in descending order for the seven most prevalent topics of the sub-database *Contrats*, estimated by the CTM. The topic numbers have been changed for illustration purposes.
Figure 1: The figure shows estimated kernel densities for cosine similarities between words created by the tf-idf/topic models (LDA and CTM) and words by the Gaius database. The number of words for the tf-idf and the number of words/number of topics of the topic models have been chosen in such a way that the standard deviations of the cosine similarities have been minimal.

Figure 2: The figure shows estimated kernel densities for cosine similarities between words created by the tf-idf/topic models (LDA and CTM) and words by the Gaius database. The number of words for the tf-idf and the number of words/number of topics of the topic models have been chosen in such a way that the average of the cosine similarities is highest.
Apart from computing cosine similarities between TM (CTM and LDA) keywording and the Gaius indexing, we have also computed similarities between the TM terms and the intuitive, non-structured SOQUIJ indexing. For the sub-database Admin, the results indicate that the TM terms coincide, on average, more with Gaius than with SOQUIJ. For the sub-databases Contrats and Travail, however, the results are reverse: the terms from the TM have higher similarities with SOQUIJ than with Gaius. Yet, the differences are minor.

3.2 Which facets and which words drive similarities

Having shown that words generated by TM correlate with human subject indexing, we are also interested in which facets mostly correlate with the automated terms. To do so, we have computed, again, the cosine similarities on the basis of five of the six facets (for the rest as described above), but leaving out successively a different one of the six facets. Deleting the terms from facet 5 (Legal category) predominantly caused the largest drop in similarities, indicating that the terms from this facet are the most important drivers for similarities. We have verified that this finding is not caused by the fact that the facets contain different relative and overall numbers of terms. Investigating in further detail which terms from which facets are the most important drivers for the similarities remains a subject for future research, especially when considering using TM for semi-supervised indexing.

3.3 Qualitative results

Above we have used quantitative methods to investigate similarities between TM keywords and human (faceted) indexing. Another, qualitative question is whether TM is capable of providing the legally interesting core of the indexed decisions and how it performs in comparison with the Gaius and SOQUIJ indexing. The first impression, which is based on our own legal expertise and a selected sample of 25 decisions, is that the quality of the automatic TM keywording as an indicator for the decision content is – with fl from decision to decision – overall encouraging. This result is certainly subjective and needs empirical validation by, for example, conducting expert and user tests.

In all three sub-databases, the vast majority of decisions (Contrats: 0.87; Admin: 0.96; Travail: 0.89) were assigned to topics, at least one of which was relevant for the decision in question with a probability of > 0.5, and a considerable proportion of the decisions (Contrats: 0.36; Admin: 0.61; Travail: 0.43) even to topics with a relevance of > 0.9, while the probability for the other two topics was much lower (mostly between 0.0 - 0.2). For those decisions where even the probability of the most relevant topic was < 0.5, there was usually another topic of considerable relevance (> 0.3).

The legal information content of the (20 or 50) topics of the three tested sub-databases was mixed. There were topics whose most probable five or ten terms associated certain facts with certain types of legal issues (certain types of cases) and others which had a rather low distinctiveness. An example of a more meaningful keywording from the Contrats sub-database is topic 2, shown in extracts in table 1 above. The word être has passed through our filter, because it was not included in our applied stop word dictionary. Topic 7 from the same sub-database, also shown in extracts in table 1, is an example of a rather meaningless topic with a large portion of irrelevant words (être, droit, peut and comme). For our qualitative analysis, we have limited ourselves to those decisions.
where at least one topic, with usable distinctiveness in the aforementioned sense, has been assigned to with a minimum probability of 0.3. In the future, it should be examined whether it is useful to manually filter the list of topics in advance. However, as our analysis has revealed, it is possible that relevant terms may originate from topics which are, taken as a whole, of little significance.

In any case, it can be determined that the keywords from those topics that were assigned to the respective decision with a high probability (> 0.8) were predominantly relevant to describe the broad outline of the decision (i.e., the type of contract, the concern). The terms, however, did not always hit exactly those points that were of particular legal interest (i.e. controversial) in the decision in question. Admittedly, that would be a high standard, which neither the Gaius indexing nor the SOQUIJ indexing have met consistently. In sum, comparing the three approaches of keywording, the Gaius indexing predominantly met best the content of the decision, which could be related to the facet scheme that pushes the indexers’ view in the direction of the legally decisive dimensions. However, Gaius does not always describe the legal core of the decision and occasionally even points in the wrong direction. This may be the consequence of the standardisation brought about by the use of a controlled vocabulary. Occasionally, the SOQUIJ indexing has been more accurate, possibly because it is not bounded by a thesaurus. TM keywording also classifies well at times as illustrated by the following example on the decision Robert c. Bergeron (2013 QCCQ 5859) from the Contrats sub-database. In that case, the buyer of a 24-year-old wooden house (plaintiff) is suing the seller asking for a purchase price reduction due to the defect of water seeping through the basement fl or. Prior to the purchase, the plaintiff had inspected the property for about only two hours without consulting an expert. When he discovered water entering next spring, he removed one of the insulation boards to find that the cement was irregularly shaped. The court dismissed the claim holding that it was not a vice caché. The buyer could have identified the defect himself by careful inspection in accordance with his duty under Art. 1726 of the Civil Code of Québec (CCQ), considering the age of the house and especially since the promise of sale (promesse d’achat) had even expressly referred to the possibility of water ingress during spring. What is legally interesting about this decision is the scope of the buyer’s duty of inspection under Art. 1726 CCQ.

Our algorithm has assigned to this decision - in this order - the topics 2, 4 and 6 with the probabilities 0.998, 0.000 and 0.000 respectively, leading to a cosine similarity to the Gaius index of 0.479. The first five terms of the topics are as shown in table 1 above. Since these terms are unigrams, but legal concepts, especially in the French language, often consist of several terms, the interpretation of the topic terms requires a certain legal expertise to recognize related terms (e.g. from salariale, équité to équité salariale). If one interprets the above terms of topic 2 in this way, and cleansing them of legally irrelevant words and redundancies as well as sorting them in a meaningful way, they read as follows: vente, immeuble, eau, malfaçon, [vice] caché. The rough context of the case is thus already drawn. The aspect of the seller’s responsibility is included in the term (vice) caché and reinforced by topic 4 (garantie; (deman-dement du) prix). However, topic 4 also misleads with the terms véhicule and moteur. Yet, the algorithm puts a probability close to zero for topic 4 (compared to 0.998 for topic 2). The probabilities must therefore be taken into account when interpreting the TM indexing. The special aspect of the buyer’s obligation to inspect, which is part of
the legal regime of *vice caché*, is not directly expressed in the TM keywording; the Gaius indexing is more explicit on this point (the term *inspection* in the Action facet), but without being unambiguous. This is also true, albeit in a different way, regarding the SOQUIJ indexing (*vice apparent* as opposed to *vice caché*). Nevertheless, neither of the two indexings, Gaius and SOQUIJ, grasps fully the legal focus of the case.

The sub-database *Contrats* contains four further decisions to which the TM algorithm has assigned the same topic combination 2, 4 and 6 in the same sequence ordered by the topic probabilities. The four documents deal with the seller’s liability for defects, and interestingly, the legal focus in three of these decisions is also on the buyer’s duty to inspect the goods, identical to the example outlined above. Only in one decision, to which our algorithm assigns topic 2 with a significantly lower probability (0.317) than with the others (0.793, 0.993, 0.646), this aspect is not the court’s main focus. This example strengthens the impression that topic interpretation must take probabilities into account; it also suggests that TM keywording might be suitable for recognizing legally similar decisions. In a corresponding manner, TM enable to use quantitative metrics such as the Hellinger distance and the Kullback-Leibler divergence to find similar documents.

4.0 Conclusion and Outlook

We have shown that similarities exist between human structured subject indexing and automatically generated terms based on methods from natural language processing. The cosine similarities between human indexing and automatically generated keywords are generally higher when using the words of topic models (TM) instead of the words generated by a frequency–inverse document frequency approach. Our quantitative and qualitative results indicate that TM can be (at least) a useful tool to support human indexing in a semi-automated approach. For example, TM can provide clustered terms that can be used, in addition to a standardized vocabulary, for indexing completion. In addition, TM can useful to identify similar documents.

We propose that future research on semi-automated indexing could try to combine TM with the multifaceted approach in two directions: first one can try to optimize TM keywording by assigning the topic terms to exogenously defined facets. This approach could also be beneficial to identify similar documents. When creating a faceted vocabulary, however, it must be ensured that the majority of the faceted terms are contained in the full text and that each word of the vocabulary is only represented within one facet. Second, it would be interesting to check whether the quality of the topics can be improved when using a vocabulary of predefined facets for the TM estimation.

References


Knowledge Organization Requirements in LIS Graduate Programs

Abstract:
The importance of knowledge and information organization has become more evident in the larger and fast changing information environment. Educating knowledge organization professionals has been a long tradition of library and information science programs. This study reports preliminary findings on knowledge organization requirements and offerings of North American ALA-accredited programs and aligns them to professional competencies. Future research and work are needed on knowledge organization competencies and curricula for the wider knowledge organization professions.

1.0 Introduction
Knowledge organization is of great importance in a global information landscape, and affects all economic, political, social, and educational sectors. Knowledge organization enhances scholarly discovery, technological innovation, entrepreneurial activity, artistic exploration and even public policy improvement (Szostak 2014). Standardization, representation, categorization, classification, description, and access all require knowledge of principles, theory, and different international and domain-specific standards and best practices. The creation and sharing of such information also require knowledge of tools and structures for storing, exchanging, and accessing this information. Preparing information professionals for careers in knowledge organization in a heterogeneous information environment is an enormous challenge for any library and information science (LIS) program, and even more challenging when most programs have limited opportunities for relevant coursework.

Educating knowledge organization professionals does not stop with cataloguing or general and domain-specific metadata standards. To educate a well-informed professional, one needs to cover the broader context of knowledge representation, classification and indexing, conceptual models, best practices, data processing, data management, information systems, and technologies, including semantic web technologies. A balance of the theoretical and practical is necessary. In addition, other topics, such as ethical considerations, intercultural awareness, responsiveness to community needs, global trends, collaboration, and resource sharing are essential for students to learn.

Drawing from an analysis of competencies developed by professional organizations, and relevant literature, this paper examines the program requirements for information or knowledge organization among the American Library Association (ALA) accredited master’s programs in North America. In addition, the paper provides a preliminary examination of their knowledge organization offerings beyond the core (required) coursework.

2.0 Competencies and Skills
What is meant by competencies and skills? Merriam-Webster dictionary defines competence as the “possession of sufficient knowledge or skills”¹. Hoffman (1999)

found that there are typically two types of definitions of competency; one referring to outputs of training as in competent performance, and the other as inputs or inherent attributes required by an individual to be able to achieve competent performance. Competencies are also defined as the “measurable or observable knowledge, skills, abilities, and behaviors (KSABs) critical to successful job performance” and are grouped into knowledge competencies (practical and theoretical understanding), skill and ability competencies (learned or natural capacities), and behavioral competencies (conduct) (Washington State 2012).

Within the LIS field, competencies are also defined as characteristics of individuals. Fisher (2001) defined three types of competencies: professional, which are occupational-related knowledge and skills required for success at a particular work setting; personal, including attributes and behaviors of individual that would apply to any setting; and educational, acquired from studying a body of knowledge in a particular field. Dole, Murych, and Liebst (2005) define competencies as a “specific range of skills, abilities, or knowledge that enable or qualify someone to perform a particular function or to carry out selected responsibilities” (125) and exclude personal and behavioral characteristics.

Several publications and presentations have examined knowledge and skills, including soft skills for metadata and cataloguing librarians but the literature on the competencies and skills required for the broader knowledge organization field is largely lacking.

Evans et al. (2018) describe the process of developing the ALCTS core competencies for cataloging and metadata professionals. They conclude that the resulting competencies document “could serve as a starting point for students, practitioners, educators, and managers to plan for an individual’s growth and development across the span of a working life, from novice to mid-career professional and beyond” (Evans et al. 2018, 195)

Several publications and presentations have addressed the skills required by cataloging professionals in library settings. They identify knowledge of and experience working with particular standards or particular knowledge organization systems, technology savviness, including semantic web technologies, metadata transformations, and metadata and ontology development skills (Joudrey and McGinnis 2014; Mitchell 2013; Carlyle 2013; Panchyshyn 2015). Fewer publications specifically address the importance of soft skills, such as effective communication, flexibility, open-mindedness, and critical thinking (Bothman 2014; Han and Hwse 2010; Hall-Ellis 2015).

Professional organizations and other groups have created competency and skills documents ranging from general lists for all librarians or information professionals to specialized lists for particular areas within the LIS field. In the North American context, ALA-accredited programs must show evidence that the curriculum “takes into account the statements of knowledge and competencies developed by relevant professional organizations” (ALA 2019, 5). Three sets of competency and skills relating to the work of library and information professionals that also address the area of knowledge organization to some degree have been used in this study: the ALA Core Competencies for Librarians, the ALCTS Core Competencies for Cataloging and Metadata Professional Librarians, and the Competency Index for the Library Field compiled by Webjunction. Due to the limited scope of this study to ALA-accredited programs in
North America, competency standards from other parts of the world have not been included here.

The ALA Core Competencies for Librarians (ALA 2009), define eight areas in which every graduate of an ALA-accredited program should know or able to do. The third area, “Organization of Recorded Knowledge and Information,” includes three subtopics of principles, skills to organize the recorded knowledge and information, and systems, standards, and methods used in this organization process.

The ALCTS Core Competencies for Cataloging and Metadata Professional Librarians (ALCTS 2017) have a more specific focus on competencies and skills for catalogers and metadata professionals within the library environment. The listing is grouped into knowledge, skill and ability, and behavioral competencies. The “knowledge competencies” cover principles, systems and technologies, and trends in the specific areas. Skills and abilities cover application of principles and standards, application of universal standards within the local context, and management of metadata within a bibliographic system. Behavioral competencies include a number of soft skills, such as communication, orientation for public service, adaptability, and problem solving abilities.

WebJunction’s Competency Index for the Library Field (Gutsche and Hough 2014) drew from a number of existing competency sets also limited to the library setting. It defines essential competencies and a number of competencies for four specific areas or functions of a library. These include competencies related to library collections, library management, public services, and technology (systems and IT). It is within the library collection competencies that “cataloging” is included, as one of six subsets. The cataloging competencies cover use of bibliographic control standards to catalog resources and management of the catalog for users’ optimal access to the collection.

Although information ethics are applicable to knowledge organization, separate codes of ethics, like IFLA Code of Ethics for Librarians and other Information Workers (IFLA 2016) and the ALISE Position Statement on Information Ethics in LIS Education (ALISE 2008) are not examined here due to their more general scope.

3.0 Knowledge Organization Education

A number of publications surveyed the current state of cataloging education in the United States. In a series of four publications of a longitudinal study, Joudrey and others examine the evolution and current state of cataloging education in ALA-accredited programs, including teaching subject analysis and subject cataloging, suggesting that the complexity of subject access required a complex approach to teaching these topics and reviewing the information organization curriculum in ALA-accredited LIS programs, courses offering, program requirements, and trends and developments in information organization education (Joudrey 2002; 2008; Taylor and Joudrey 2002; Joudrey and McGinnis 2014).

Saye (2002) addressed cataloging as a required course, curricular changes to include new areas within cataloging, the influence of LIS faculty on cataloging courses within LIS education, and the changing perception of the importance of cataloging in library and information studies programs. Davis (2008) provided an overview of the state of cataloging courses in ALA-accredited programs at the time, concluding that while most programs are offering courses on cataloging, and even requiring them, they tend to rely on introductory courses for the bulk of cataloging education.
In a series of publications, Hudon (2011; 2014) examined the knowledge organization and bibliographic classification instruction in ALA-accredited LIS programs in Canada and the United States, taking a closer look at course objectives. Among the findings are that teaching and learning objectives “are very general, and often rather vague, covering in a single statement many concepts and subjects” (Hudon 2014, 530).

The balance of theory and practice in cataloging courses is a frequent topic of discussion. Many surveys of cataloging education considered the debate as one of the factors shaping curriculum, but several publications addressed the balance more directly (Moulaison 2012; Normore 2012; Intner 2002). Perceptions of students and supervisors on practica are also discussed. Damasco and McGurr (2008, 2010) conducted surveys of entry-level cataloguers and practicum students and supervisors to identify attitudes about practica and best practices for the practicum as part of the LIS curriculum. Interviews of LIS graduates conducted by Snow and Hoffman (2015) found that balancing theory and practice, teaching hands-on cataloging practice, and placing cataloging in a real-world context contribute to effective learning.

4.0 Methodology

The scope of this study is limited to ALA-accredited programs in North America. Two programs were excluded from the directory provided by the ALA Office on Accreditation2. One due to the fact that curriculum information was not available in English and the second because it was granted accreditation around the data collection period. A total of sixty-four (64) accredited programs housed in sixty (60) academic units were included in this study. Among these academic units, 43 (67%) have some membership in the iSchools organization (ischools.org) and 21 (33%) do not have any type of iSchools membership.

For each program, the following data was collected: 1. required (core) information or knowledge organization coursework, defined a course required by all students in the accredited program, regardless of specialization path; 2. whether the program offers a set of courses that allow students to specialize in the area of information and knowledge organization; and 3. courses included in the specializations. For each course, the title and course descriptions were included in the analysis, which employed content analysis techniques using both manual coding and use of NVivo for the emergence of major themes.

Competencies and skills relevant to knowledge organization from three sets were examined; the ALA Core Competencies for Librarians, ALCTS Core Competencies for Cataloging and Metadata Professional Librarians, and WebJunction’s Competency Index for the Library Field. Competency statements were then aligned to the course content as described in the official course description provided by each program on their websites or university catalogs of courses.

5.0 ALA-Accredited Programs

Of the sixty-four programs examined in this study, 33 (52%) offer a Master degree of Library Science/Studies (MLS) or Library and Information Science/Studies (MLIS), 19 (30%) offer a Master of Science (MS) in LIS, information, or information science, and

2 http://www.ala.org/educationcareers/accreditedprograms/directory
12 (19%) offer a Masters of Arts (MA), Information (MI), Information Systems (MIS), Management (MM), etc.

Of the sixty-four programs, 60 (94%) have a core course in IO/KO and 31 (51%) have a defined area of study in IO/KO identified either as a concentration, specialization, cluster, or a pathway. This is not to say that programs that do not declare a defined area in IO/KO do not offer additional coursework in this area.

When comparing membership in iSchools, of the members, 20 (47%) offer MLIS degrees, 14 (33%) offer MS, and 8 (19%) other Master degrees, with 38 (88%) requiring an IO core course, and 24 (56%) offering a specialization in the area. Of the non-iSchools, 13 (62%) (19%) offer MLIS degrees, 4 (19%) offer MS, 4 offer other Master’s degrees, with all (100%) requiring a core course in IO/KO, and 7 (33%) offering a specialization in IO/KO.

6.0 Required Core Coursework in Knowledge Organization

The majority of the programs, 94%, have a requirement for a course that covers topics in knowledge organization. Most of these courses are required of all students in the program. A few programs give an option between a select list of courses, one of which is an knowledge organization course.

Required core courses in the IO/KO area were normalized and analyzed. Figure 1 shows title terms that appear in at least two core course titles. A comparison of the use of information vs. knowledge shows that “information” is used in forty-four core course titles, where “knowledge” is used only in thirteen course titles. Three courses have both information and knowledge in their title. Other terms appearing in the titles include “access,” “representation,” and “description.”

Frequency of terms appearing ten or more times in course descriptions, are presented in Figure 2. The terms were normalized for reporting. The term “systems” is one of the most frequent term, with a frequency of forty, followed by the terms “theory” and “practice,” each appearing thirty times. Standards is also one of the top terms (frequency of 27). Concepts, structures and tools have the same frequency (19), followed by subject and vocabularies. Some descriptions mention specific standards and tools, such as the Library of Congress Subject Headings (LCSH), Dublin Core, Resource Description and Access (RDA), MARC, BIBFRAME, and XML, among others.
It is interesting to note, that although the core courses are introductory and content coverage is targeting all students in the program, preparing them for all types of IO/KO study, the term “library” appears fifteen times in these descriptions. This may be an indication that introductory IO/KO courses do not necessary cover knowledge organization in all settings. A more detailed examination of course content and learning outcomes is necessary to shed light on this observation.

![Figure 2: Core IO/KO Description Terms](image)

Description statements were coded using the NVivo software for thematic analysis, which resulted in a number of content themes emerging from this analysis. The top themes, appearing in ten or more course descriptions, are presented in Figure 3. The theme, a description of what is included in each theme when necessary, its frequency, and alignment to the three competency sets, ALA Core, ALCTS Core, and WebJunction’s competencies are included in the respective columns of Figure 3.

The top three themes with a frequency higher than twenty are “principles,” including theories of IO, “information,” and “standards.” In this thematic analysis of the content, the theoretical coverage occurred more frequency than the practical aspects of IO/KO. Library cataloging received equal treatment as metadata creation. The two combined (freq. of 20) approximate the frequency of classification (freq. of 18). Very few (3) described the coverage of IO in broader contexts to address national and international coverage but also the need to implement universal (international) standards to local contexts for more effective community service. Settings, on the other hand, either specified (e.g., libraries), or inclusive (e.g., “in different environments”) received more attention in the core courses.

Three themes have a direct alignment to all three competency sets: classification, cataloging, and context, although the broader context is not a topic that appears in many course descriptions. This can be explained by the scope of the competency sets, which are limited to libraries. In addition to these two topical themes, four themes, principles, standards, metadata, and IO approaches, align to both ALA and ALCTS Core
competencies. Fourteen themes align to the ALA Core statements, twenty-three to ALCTS Core statements, and nineteen align to WebJunction’s statements. The hunger number of alignments in the latter two can be explained by the detailed competency statements included in them vs. the very broad statements of ALA Core.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Description</th>
<th>Frequency</th>
<th>ALA</th>
<th>ALCTS</th>
<th>Webjunction</th>
</tr>
</thead>
<tbody>
<tr>
<td>principles</td>
<td>theory, theoretical, principles</td>
<td>39</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>information</td>
<td>what is organized</td>
<td>31</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>standards</td>
<td>description, metadata, encoding</td>
<td>23</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>access to information</td>
<td>to information, resources, collections</td>
<td>19</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>retrieval-discovery</td>
<td>theory and methods, retrieval, discovery</td>
<td>19</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>practice</td>
<td>practice, practical</td>
<td>19</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>settings</td>
<td>environments where IO happens</td>
<td>19</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>classification</td>
<td>classification, categorization</td>
<td>18</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>resource description</td>
<td>more inclusive than cataloging or metadata</td>
<td>12</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>retrieval systems</td>
<td>retrieval systems, search engines, discovery systems, catalogs, indexes, finding aids</td>
<td>12</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>subject analysis</td>
<td></td>
<td>11</td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>systems &amp; technologies</td>
<td>technologies, database structures, information systems, development of information systems, platforms,</td>
<td>11</td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>knowledge</td>
<td>what is organized</td>
<td>10</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>information structures</td>
<td></td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>metadata</td>
<td>creation, selection of elements</td>
<td>10</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>cataloging</td>
<td>library cataloging</td>
<td>10</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>representation</td>
<td>representation of information</td>
<td>9</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>IO approaches</td>
<td>methods, techniques, approaches to IO</td>
<td>9</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>authority control</td>
<td></td>
<td>7</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>impact of IO</td>
<td>impact on access, retrieval, users</td>
<td>6</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>searching</td>
<td></td>
<td>6</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>IO systems development</td>
<td></td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KOS types</td>
<td>examination of KOS: controlled vocabularies, classification schemas, subject headings, taxonomies, ontologies</td>
<td>5</td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>resource types</td>
<td>types of materials, formats</td>
<td>5</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>assessment</td>
<td>evaluation, assess metadata quality, accuracy, assess tools, systems</td>
<td>4</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>application of standards &amp; principles</td>
<td>specifying knowledge of and ability to apply principles</td>
<td>4</td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>bibliographic networks</td>
<td>bibliographic databases, cooperative bibliographic databases</td>
<td>3</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>context (local, national, international)</td>
<td>local context; broader context (national, international)</td>
<td>3</td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>human information behavior</td>
<td>broader theories and models than searching</td>
<td>3</td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>indexing</td>
<td></td>
<td>3</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>vocabulary creation</td>
<td></td>
<td>2</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>data</td>
<td>data management, organize data</td>
<td>2</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>related fields</td>
<td>cognitive sciences, semiotics, artificial intelligence, philosophy, linguistics</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>trends</td>
<td>trend awareness</td>
<td>2</td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>conceptual models</td>
<td></td>
<td>1</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>linked data</td>
<td></td>
<td>1</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>professional organizations</td>
<td></td>
<td>1</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>user needs</td>
<td>addressing user needs, vs access by users</td>
<td>1</td>
<td>x</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3: IO/KO Core Course Themes and Competency Alignment
7.0 Knowledge Organization Specializations and additional Coursework

Only about half of the programs offer a set of courses that is either identified as a cluster of courses, pathway, specialization, or concentration in knowledge organization, which allows students to have a more in-depth exposure in this area, beyond their core requirements. The data show that a higher percentage of iSchool programs offer a specialized program of study in IO than non-iSchool programs. An examination of the names (normalized) used for these specialized areas of studies, indicate that by far, “information organization” is the most commonly used name:

<table>
<thead>
<tr>
<th>Specialization</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information Organization</td>
<td>18</td>
</tr>
<tr>
<td>Knowledge organization</td>
<td>2</td>
</tr>
<tr>
<td>Technical Services</td>
<td>6</td>
</tr>
<tr>
<td>Cataloging</td>
<td>2</td>
</tr>
<tr>
<td>Library studies</td>
<td>1</td>
</tr>
<tr>
<td>Collection management</td>
<td>1</td>
</tr>
<tr>
<td>Information Services, Organization, Management, and Use</td>
<td>1</td>
</tr>
</tbody>
</table>

The number of relevant courses associated with these specializations, ranging from none to sixteen distinct titles. In the cases where no courses are listed within the area of specialization, it is possible that a set of electives is available for the students to taken in consultation with their advisors. Courses listed in these specializations cover a variety of topics, as illustrated by Figure 4, with the most frequent being cataloging and classification, metadata, archives and records management, and KOS construction.

<table>
<thead>
<tr>
<th>Courses</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cataloging &amp; Classification</td>
<td>24</td>
</tr>
<tr>
<td>Metadata</td>
<td>15</td>
</tr>
<tr>
<td>Archival Description, Records Management</td>
<td>10</td>
</tr>
<tr>
<td>KOS/Thesaurus Construction</td>
<td>10</td>
</tr>
<tr>
<td>Classification, Subject analysis</td>
<td>8</td>
</tr>
<tr>
<td>Indexing (&amp; Abstracting)</td>
<td>7</td>
</tr>
<tr>
<td>Technologies</td>
<td>6</td>
</tr>
<tr>
<td>Advanced Cataloging &amp; Classification</td>
<td>3</td>
</tr>
<tr>
<td>Collections</td>
<td>3</td>
</tr>
<tr>
<td>Databases</td>
<td>3</td>
</tr>
<tr>
<td>Digital Libraries</td>
<td>3</td>
</tr>
<tr>
<td>Information Architecture</td>
<td>3</td>
</tr>
<tr>
<td>Information/Data Management</td>
<td>3</td>
</tr>
<tr>
<td>Advanced Metadata</td>
<td>2</td>
</tr>
<tr>
<td>Linked Data</td>
<td>2</td>
</tr>
<tr>
<td>Knowledge Management</td>
<td>2</td>
</tr>
<tr>
<td>Visualization &amp; Visual Analytics</td>
<td>1</td>
</tr>
<tr>
<td>Legal issues</td>
<td>1</td>
</tr>
<tr>
<td>User Experience Research</td>
<td>1</td>
</tr>
<tr>
<td>Big Data Analytics</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 4: Specialization Courses

In most cases, additional courses are available to the students to further explore allied areas, such as rare book librarianship, digital technologies, text mining, and digital preservation. Additional data are needed for a more in-depth analysis of the content of additional elective courses and the frequency they are offered. During the data collection process, it became evident that many courses listed on the course catalog have not been offered for a while. In addition, many programs are revising their curriculum (a few program requirement and offerings changed at the start of the 2019-2020 academic year), which could be the result of continuous curriculum review, reorganization of academic units, trends, and shift in direction, all of which may affect the focus of the programs or the availability of courses.

The literature points out the importance of a balance between theory and practice and the curriculum option for practica as part of LIS education. The data show that overall, the majority of the programs (78%) provide an option or a requirement for an internship or practicum to be counted towards the degree. Among programs that offer an IO/KO specialization, eight (8) require an internship or practicum as part of the program, two
(2) offer it as one of the options of an end-of-the-program requirement, sixteen (16) offer it as an option, 4 require it or highly recommend it to particular groups of students, and seven (7) do not offer it as an option or do not provide information about internships or practica. Among the programs that do not list an IO/KO specialization, two (2) have it as a requirement, six (6) offer it as one of the options of a program requirement, eleven (11) offer it as an option, one (1) require it only for a particular group of students, and six (6) do not offer an option or do not provide information on internships.

8.0 Conclusion
In summary, this preliminary examination of the knowledge organization required courses and specializations in ALA-accredited programs shows that the majority of programs require an introductory course in this area. Not all programs offer courses beyond the required course, which means that students have a limited exposure to IO. This may be of concern, considering the important role knowledge organization plays in today’s information landscape, which goes beyond libraries.

Additional research is needed to examine the content of the courses in more detail, the learning objectives, and their applicability to knowledge organization professionals in any information environment. Extending the scope beyond the ALA-accredited programs with provide insight on the needs for educating and preparing knowledge organization professionals world-wide.

In addition, there are no international competencies, including knowledge and skills, for the broader field of information and knowledge organization, regardless of setting. Is it perhaps the responsibility of an international professional organization such as IFLA or ISKO to undertake such a task? How can LIS programs benefit from an international effort to define IO/KO competencies to adequate prepare their graduates for careers in this field?

References


Saye, Jerry D. 2002 “Where Are We and How Did We Get Here? Or, the Changing Place of Cataloging in the Library and Information Science Curriculum: Causes and Consequences.” Cataloging and Classification Quarterly 34, nos. 1–2: 119-141.


Knowledge Organization (KO) has a historically established construction based on a concept theory according to the Aristotelian model. It is from the Organon that the primary theoretical basis of documentary-language thinking is established. One can identify the construction of conceptual thinking in classical approaches such as Ranganathan's classification theory. More directly, Dahlberg's theory of concept proves the demarcating point of view of the Aristotelian foundation in KO. However, even in such approaches as Ranganathan's, one can also identify another theoretical tradition in classification theory concerned with constructing a non-conceptual approach. It is from Emanuele Tesauro that we can conceive a non-conceptual theory, establishing a given conceit theory. This research aims to point out the dichotomies and correlations between a concept theory and a non-conceptual one in KO. The research method is theoretical, structured from the perspective of a historical epistemology with a pragmatic background. At first, the historical-theoretical place of concept theory in the KO is determined. Later, the foundation of a non-conceptual theory (or conceit theory) is identified. Furthermore, the relationship between theories in historical and contemporary development in the classification theory is discussed. From the Aristotelian categories, one can understand a method for the relationship between signifier, meaning, and referent, and establish the associations of meaning between terms. The non-conceptuality theory can be historically identified in Emanuele Tesauro. From his Categorical Index, published in the Cannocchiale aristotelico, the condition of concept (and no-concept) is the apex of a variety of chain of rhetorical being, and it points to the creation and progressive unfolding of language figures (conceit elements). According to the French philosopher, Barbara Cassin, the theory of non-conceptuality lays the foundation or logology perspective for a semiotic-semiological-pragmatic focus and encounters the later Wittgenstein, Michel Foucault, and symbolic and poststructuralist approaches. Coincidentally, sources such as Bernd Frohmann, Søren Brier, and Hope Olson will devote themselves to the study of these philosophers to re-discuss the KO. The results lead us to the discussion of different ontological distinctions of language in the tradition of research in KO, as well as the possibilities of critical-social construction of a non-conceptual perspective. The conclusions point to the common origin of the Aristotelian nature of conceptual thinking and non-conceptual thinking.

1.0 Introduction

Knowledge Organization (KO) has a historically established construction based on a concept theory according to the Aristotelian model (Aristotle 2010). It is from the Organon that the primary theoretical basis of documentary-language thinking is established - as well as from reasoning about the role of language in knowledge, already present in Plato's Cratylus (Plato 1963). One can identify the construction of the conceptual thinking in classical approaches such as Ranganathan's classification theory (Saldanha 2014), despite the possibility of pragmatic openness launched by the Indian mathematician.

More directly, Dahlberg's theory of concept establishes the demarcating point of view of the Aristotelian foundation in KO (Dahlberg 1993; 2006). However, even in such approaches as Ranganathan's (Ranganathan 1967), one can also identify another theoretical tradition in classification theory concerned with constructing a non-conceptual approach. It is from Emanuele Tesauro that we can conceive a non-conceptual theory, establishing a given conceit theory.
According to the review by Hjørland (2008), the theoretical path investigated here may point to the distinction in KO between a positivist tradition, focused on hierarchies already pre-existing in nature, and a pragmatic one towards the social construction of local needs and interests. We will try to demonstrate how the second approach has been gaining since the 17th century, the opening to pragmatic-critical thinking in KO.

This research aims to point out the dichotomies and correlations between a concept theory and a non-conceptual theory in KO. The research method is theoretical, structured from the perspective of a historical epistemology with a pragmatic background. The research is structured in order to configure the role of the philosophy of language in the construction of meta-informational practices - as observed by Blair (1992). The final focus of the reflection seeks to demonstrate how, since Emanuele Tesauro's proposals, we can anticipate the opening to semiotic, symbolic, and critical approaches to KO.

The research represents a development of Brazilian studies in historical epistemology in KO. Specifically, it is the result of a set of investigations that have been carried out based on the work and thought of Emanuele Tesauro. The present study is, in this way, the result of other research already published, namely, Saldanha, Almeida, and Silva (2020), Saldanha et al. (2018), Saldanha and Souza (2017), Saldanha (2017a; 2017b).

2.0 The discursive path: between ontology and logology

From the Aristotelian categories, one can understand a method for the relationship between signifier, meaning, and referent, and establish the relationships of meaning between terms. In KO, the non-conceptuality theory can be historically identified in Emanuele Tesauro. From his Categorical Index, published in the Cannocchiale aristotelico, the condition of concept (and no-concept) is the apex of a kind of chain of rhetorical being, and it points to the creation and progressive unfolding of language figures (conceit elements).

The non-conceptuality theory has been synthesized directly from Hans Blumenberg's theory, Barbara Cassin's philosophy, and the rhetorical approach of the Belgian collective “Groupe μ” (1970). Openness to non-conceptuality, from the empirical point of view, represents the recognition of plural horizons of meaning. Therefore, the theoretical and applied approaches that address, for example, discursive issues of gender, race, and social classes, such as those made possible by the researches of Hope Olson (2003) and Antonio García Gutiérrez (2011). On the one hand, the issues of feminism in the context of knowledge representations, on the other, the need for a tropological theory of classification, demonstrate the already advanced application of discursive theories in KO.

For Blumenberg (2017), it is necessary to consider that concepts are not only based on objects, but they also constitute objects. Concept, therefore, can be defined as a rule of representing representations - therefore, a representation of a representation. The hypothesis of the existence of the concept presupposes, therefore, the urgency of non-conceptuality. This means that we are not in a logical-essential terrain of the object as we are talking of representations of representations.

The concept theory leads us to a rigorous semantics capable of establishing an ontology that sustains being precisely by being itself, a classical conception that finds in Heidegger its simulacrum in language and the so-called first Wittgenstein (1979). The
theory of non-conceptuality lays the foundation – or logology perspective, according to the French philosopher Barbara Cassin (2005) - for a semiotic-semiological-pragmatic focus (as the research by Friedman and Thellefsen (2010)) and finds in the later Wittgenstein (1979), Michel Foucault (1971; 2002), and symbolic and poststructuralist approaches, including revisions in Information Science, as the work of Day (2005).

According to Cassin (2005), ontology operates in the realm of being, the discourse that manifests and affirms being. At the ontological level, the discourse imposes itself and imposes what can be said. Logology acts in the discourse that makes being, that is, being is an effect of saying. In the logological dimension, discourse produces what is said. The ontological movement depends, therefore, on the concept as a way of sustaining its imposition.

Fundamental to the establishment of metaphysics in Western culture, a concept theory will also be central to the late foundation of KO. This theory supports, by the principle of identity, the logical possibility of subjects and machines to sustain formal relations of meaning. However, they are not configured as valid from a socio-cultural point of view, due to the dynamics of representations. These dynamics are precisely the place where metaphor lives: in a non-conceptual world. In other words, we are here in the political world. This is the central point of social life observed by Cassin (2005) from the linguistic-rhetorical construction of the real since Ancient Greece: before logic, existence is logological, the existence in the city, in the gregarious world of the subjects, happens through the intensive use of non-logical language. This is the opening for political thinking about language. Its repercussion in KO is direct; it represents criticism of pure logic in classification practices.

In Information Science and KO, we can also find these tendencies as the focus of pragmatic theory and critical theory for informational practices in the late 20th century. One of the central syntheses of these approaches is in the discussion about the hermeneutic-rhetorical paradigm of the informational field observed by Rafael Capurro (1992). Equally, Bernd Frohmann (1990), Søren Brier (1996), and Hope Olson (1999; 2003), for example, devote themselves to the study of these philosophers to re-discuss the KO. The results lead us to the discussion of different ontological distinctions of language in the tradition of research in KO, as well as the possibilities of critical-social construction of a non-conceptual perspective, from a pragmatic-critical approach.

3.0 “The courage of metaphor:” challenges for KO from Emanuele Tesauro

If rhetoric was relatively critical during the Middle Age, we could see it became a trending topic during the Renaissance and, after the Council of Trento, it certainly acquired a special place in western culture.

One of the main interests in studying the work of Tesauro concerns the possible uses of metaphors. Umberto Eco, who widely wrote on metaphors, rooted his reflections in the Aristotelian conception of metaphor as “instrument of knowledge, clearness, and enigma” (Eco 1984, 164), stating that its function is to show, teach to see and recognize similarity among things, “the subtle web of proportion between the cultural unity” (Eco 1984, 164).

As we will see, Tesauro's work is an encyclopedic model in which we can recognize the core of Aristotelian rhetoric, establishing the metaphor model as a way of discovering unedited relations between knowledge data.
Aristoteles considers metaphors, mainly when they are subtle and not evident, as learning figures: “the obvious metaphor, the one that does not shock at all, is rejected. When metaphor shows us things opposite to what we used to believe, it becomes clear that we learned, and it looks like our mind is telling us: ‘it was like this, and I was wrong’” (Eco 1984, 164).

Metaphor is capable of “putting in front of the eyes” an unexpected relation between things, it “imposes a reorganization of our knowledge and opinions” (Eco 2007, 70), becoming a “productive tool of new ontologies” (Eco 2007, 67), capable of leaving “a trace in our encyclopedia” (Eco 2007, 72).

Eco states that during the Middle Age, an idea of metaphor developed in which its sense shifted away from the Aristotelian definition, acquiring a purely ornamental function, with very little knowledge. (Eco 2007).

According to Blumenberg (2017), in his theory of non-conceptuality, metaphor is initially a disturbance of connections, a fracture in the homogeneity that allows mechanical reading. In this way, metaphor is historically understood as a semantic anomaly. This historical reason places the priority of the concept over the metaphor, as being the first “correct,” and the second a displacement of meaning. However, it is precisely in the analysis of the most straightforward and most complex concepts, like the cosmos, that metaphor is revealed as, precisely, essential for knowledge (before and after the concept).

Cassin (2005) demonstrates that, through the reinterpretation of Aristotle's Poetics, that metaphor has already produced a “better teaching” than logic, since they produce a picture of reality. Metaphors put before the eyes things they mean in action.

Eco compares Middle Age and Renaissance proposals on the use of metaphors to the theory developed in the 17th century by Emanuele Tesauro, observing that during the Middle Age, the Aristotelian theory was not much prized. Simultaneously, Tesauri’s Il Canocchiale Aristotelico (1654) represents its prime reach (Eco 2007, 123-124).

Tesauro seems to be Aristotle's heir, mainly when it comes to the knowledge function of metaphor. All the possible metaphorical combinations that Tesauro suggests and lists along with his quite broad “categorical index,” considered by Eco as a “model of an unlimited semiosis,” or encyclopedic web of interpreters, allow discovering “a warehouse of unedited metaphors” (Eco 1984, 170).

Eco's starting point is the Aristotelian classification of different kinds of metaphors to develop his analysis of production and interpretation. The first two, from gender to species and from species to gender, can be classified as synecdoche, while the third one, from species to species, should be thus considered a metaphor (Eco 1984, 150-154).

The first two kinds of metaphor may be satisfactory when described through the dictionary model, which is hierarchic, designed as a tree, where we can find the neo- platonic heritage descending from the so-called “Porphyrian tree.” It manages metaphors in the kind of relation from gender to species and vice-versa, which means synecdoche through the formula pars pro toto or totum pro parte (Eco 1990, 146-147).

In order to describe the third and the fourth kind, that is metaphor stricto sensu, we need to resort the non-hierarchic model of an encyclopedia, a web or rhizomatic structure in which to insert the practices of interpretation and of knowledge of the so-called “unlimited semiosis,” where metaphor has the function of an interpreter. Inside the sophisticated space of an encyclopedia, metaphor is a constitutive part of a web of
interpreters: “of a reticulate of properties in which parts are the interpreters of others” (Eco 1984, 177), properties furthermore sensitive to circumstances, historical contexts and knowledge of the world, fact that turn it “interesting.”

The substantial opposition between the two models of dictionary and encyclopedia is based on their properties, being analytics in the first model and synthetic in the second one. As Eco states (1980, 206), the properties of dictionaries are “artifices of a metalinguistic economy that allows avoiding a long list of encyclopedic properties.”

The content of the system designed in Tesauro's Cannocchiale is organized encyclopedically, revealing that the metaphorical shifts are guaranteed by the language's very structure rather than by ontological relations.

Tesauro studies the refinement in single words and short texts as well as in visual and plastic manifestations: for him, mastering metaphors is not casual but requires practice.

Step one of the path leading to master the craft and use of metaphors consists of reading catalogs and all varieties of collection of texts and emblems., in order to learn how to combine them.

Step two is drafting a categorical index (both on cards and tables). This index has the task of modeling a universe “semantically organized” (Eco 1980, 214). Branches, collecting everything that the ten Aristotelian categories can include, are established, and things can be found just sliding through the index.

Tesauro follows his time trend, testifying the encyclopedist's effort to flee from a dry classification based on synecdoche (from gender to specie and vice-versa). He proposes a partial accumulation (partial, because at least it is ordered by the ten Aristotelian categories of Substance plus the nine accidents) that allows the inventio, in its meaning of discovering, of unexpected relations between the “objects” of knowledge.

According to Proctor (1973), Tesauro conceives a figurative approach or a structure that addresses the relationship between sense, emotion, and intelligence. His experiments seek not only to isolate and relate terms to identify their sign and symbolic effects. The theoretical Tesaurian activity also allows, in its combinatorial technique, to identify other meanings based on the relationship between ideas.

In his work “A Theory of the Conceit,” Proctor (1973) demonstrates that conceit represents a particular image of comparison, which produces different figures of speech, including metaphor. These figures are used in the singularities of objects, in their complexity and tension between objects. In Proctor's (1973) point of view, Tesauro solves historical questions of the idea of figurative language studies.

Aware of the difference between science and rhetoric, the seventeenth century, the use of metaphor as a way to discover pre-existing relationships. The methodological path of the Tesauro takes us, thus, deepening Aristotelian thought, to apprehend the object and its representation by imagination. The “imagination,” we can understand, in contemporary terms, as the socio-cultural imaginary that composes the meanings of objects and their relationships in the real.

Emanuele Tesauro (1654) visits the same Aristotle we knew from the logical tradition in KO in the 20th century – centrally, Organon, Rhetoric, and Poetics (Aristotle 1966; 1991; 2010). The difference lies in his understanding of the complexity of the Aristotelian work concerning language. It is about verifying the relations between logic and rhetoric and intensively exploring its developments. Specifically, through the metaphorical path, in the context of rhetoric, Tesauro allows us to discuss the question
of bringing the social and cultural world. The result is the repositioning of political issues at the center of reflections and practices on the organization of language and knowledge.

This movement establishes a path at the end of the twentieth century for critical theory in KO - even without talking directly with the Tesauro vanguard. We can find this critical theory, for example, in Olson (1999; 2003), in Brier (1996) and García Gutierrez (2011). It is important to note that part of the critical solutions pointed out by these authors are already in Emanuele Tesauro, from rhetoric - It is about the possibility of perceiving a classical non-hierarchical structure, the Porphyrian structure; as well as perceiving the plurality of meanings in the socio-cultural world. In other words, as described by Capurro (1992), “The rhetorical distinctions do not intend to separate informative (and deliberative) speech from the other forms of speech nor to isolate all of them from ethics and politics.”

This is the direction towards plurality via a theory of the conceit. In non-conceptuality theory, Blumenberg (2017) indicates that poverty of our relationship with reality (amid the richness of our relationship with possibility) is not only of the order of knowledge but also of language. The universe of action (that is, the universe of pragmata) opens us up to the possibilities for transforming the social world.

The lesson of Tesauro, therefore, represents, in our analysis, what Blumenberg (2017) will call “the courage of metaphor.” This courage reveals to us the dilemmas of the economic effect of the concept (including its need, our conceptual urgency). However, the same metaphorical courage reveals to us that beyond poverty (our semantic limits) and the need for economy of meaning for communication there is our freedom.

Metaphor, the object of exploration of the theoretical Tesauro's thought, therefore represents a challenge for KO. The practices of knowledge representation gain here the horizon to face the socio-cultural dilemmas impregnated with alleged neutrality.

Between ontology and logology, Emanuele Tesauro opens, with his theoretical and applied reflections (the conceit theory), the path to the second direction in KO. As seen, logology - a term consecrated by Bárbara Cassin (2005) - operates in the territory of the discourse. This territory is marked by an understanding of the social production of knowledge and its representations. In KO, logology can be expressed by the significant influence of a pragmatic theory, already detected by Hjørland (2008). Foucaultian and Wittgensteinian contributions are registered here. The political condition of language seems to us most fundamental in the theoretical-applied window opened by Tesauro. The study of metaphor brings us to the reflection on the effects of saying, on the risks of the word, on the need to critically understand the practices of representing knowledge as a form of oppression, but also of autonomy. Therefore, the logological path in KO is an itinerary towards the socio-cultural dilemmas of a general theory of classification.

4.0 Conclusions: KO as politics

The metaphorical path of knowledge operates in the potential and pragmatic plurality of the senses – logological mode, according to Cassin (2005). The concept is achieved in the search for an abstract and watertight singularity of meaning – ontological mode. In this way, the conclusions point to the common origin of the Aristotelian nature of conceptual thinking and non-conceptual thinking. Besides, the conclusions aim to recognize each approach's uniqueness for solving the current semantic problems of
information systems, as well as for historical gaps in theoretical understanding, especially non-conceptual studies.

Tesauro (1654) has made possible, since the 17th century, a logological form of thought for the organization of language through the foundations of logic and rhetoric. From the plan applied to the political plan in KO, the logological (or discursive) perspective marks a turning point in thinking in classification. In other words, metaphor extinguishes precisely the separation between the concept and the policy, nullifies the possibility of neutrality discourse in informational practices.

We have here the locus for the deepening in KO of theoretical and methodological foundations for a semiotic theory as seen in Brier (1996), a critical theory, as developed in García Gutiérrez (2011) and a symbolic theory - for example, from Todorov's reflections, as seen in Saldanha (2017b).

Acknowledgment
This research was funded by CNPq-Brazil and Faperj-Brazil.

References
A Faceted Conceptualization of Digital Object Reuse in Digital Repositories

Abstract:
In this paper, we provide an introduction to the concept of digital object reuse and its various connotations in the context of current digital libraries, archives, and repositories. We will then propose a faceted categorization of the various types, contexts, and cases for digital object reuse in order to facilitate understanding and communication and to provide a conceptual framework for the assessment of digital object reuse by various cultural heritage and cultural memory organizations.

1.0 Introduction
The emergence of large scale digital libraries and repositories such as HathiTrust, the Internet Archive, Europeana, and the Digital Public Library of America provides new opportunities for digital information users to openly and freely interact with, use and reuse a broad range of digital objects. The unprecedented availability of massive digital collections of books, manuscripts, images, photos, and maps offers new, individual and collective ways of making sense of information and of creating digital content. The implications of using, reusing, and repurposing digital content and collections, in this open and information-rich environment, are profound and multifaceted, cutting across many different institutional contexts such as libraries, archives and museums, as well as numerous disciplines and a wide range of user communities and audience types. Access to digital libraries, in particular, has been closely associated with and discussed in regards to such measures of impact, value, and usefulness. In fact, a cursory glance at many digital library evaluation models that have been developed in the past two decades, demonstrates the importance of impact assessment and usability of digital libraries. In today’s world, digital information users, academic as well as the general public, are able to make use of digital libraries to read, explore, entertain, write, research, create, and contextualize. Users engage in a diverse range of information practices and tasks, including searching, retrieving, using, learning, conceptualizing, synthesizing, presenting and disseminating. The reuse of digital objects can, for instance, be imagined in the context of educational, cultural, artistic, historical, geographical, and genealogical research and exploration. Recent research stressed the importance of and the need for developing a digital object reuse assessment framework that would support the digital library community in understanding, identifying, and categorizing digital object reuse cases and contexts to assist in measuring the value and impact of digital objects, not only for academic communities, but also for the general public (O’Gara, 2018). Such a framework needs to provide a clear understanding of the nature, type, contexts, and
instances of digital object reuse in digital libraries, archives, and repositories. In order to provide a comprehensive perspective of what constitutes digital object reuse, this paper aims to propose a faceted conceptualization of digital object reuse cases, contexts, instances, and examples in order to facilitate the understanding and communication of the reuse of digital objects in various digital information environments for educational, research, and general purposes. An examination of the literature and the results of a survey of digital repository managers on digital object reuse are the basis for the development of a number of key conceptual facets that could be used to understand the nature and types of digital object reuse and its various facets. Examples of the facets being developed include: The Nature of Content (content type, document type, media type), Process (aggregation, derivation, mash-ups, enhancement, reformatting, synthesis, repackaging, curation), Goals/intention (Gaming, reading, reviewing, researching, combining data sets), Agents (scholars, archivists, curators, librarians, records managers, researchers, students, readers, digital repository managers), and Environment (social media, data archives, institutional repositories, digital libraries, community archives).

2.0 Prior Research
2.1 Definitions of Digital Object Reuse

Wikipedia (Wikipedia contributors, 2020) defines the term Reuse as “the action or practice of using an item, whether for its original purpose (conventional reuse) or to fulfil a different function (creative reuse or repurposing). On the other hand, reuse draws upon a nuanced understanding of how an object is repurposed or transformed”.

“In a discussion of scientific data reuse, Pasquetto et al. (2017) argues that it is challenging to distinguish between use and reuse and defines the two terms as follows: "In the simplest situation, data are collected by one individual, for a specific research project, and the first “use” is by that individual to ask a specific research question. If that same individual returns to that same dataset later, whether for the same or a later project, that usually would be considered a “use.” When that dataset is contributed to a repository, retrieved by someone else, and deployed for another project, it usually would be considered a “reuse.””

In a review of definitions for use and reuse in relation to research data, van de Sandt et al. (2019) identified four components in those definitions, namely character of data, user, purpose, and time. However, they argued that none of the identified characteristics from the discourse proved a difference between reuse and use and as a result they provided the following definition:

“Thus, we define (re)use as the use of any research resource regardless of when it is used, the purpose, the characteristics of the data and its user.”

Thompson et al. (2017) define use “as the process of accessing particular content. Often knowing that a user has “visited” or “downloaded” an object satisfies evaluation criteria for this category.

2.2 Assessment of Digital Object Reuse

Assessment of the value, impact, and usability of digital objects requires a holistic and multidimensional framework that takes into account use, reuse, and repurposing of digital objects. While user studies have contributed significantly to how we measure, assess, and evaluate digital libraries as a whole, we need to be able to demonstrate the value and impact of digital libraries at the digital object level as well. There are quantitative measures associated with the use and reuse of digital objects such as the
number of clicks, downloads, bookmarks, views, likes, and items shared and saved. However, these measures only provide a general indication of impact and merit. We need methods that document and demonstrate the quality, extent, and nature of the use and reuse of digital objects in relation to such facets as the context, discipline, collection that the object belongs to, geographic origin, time period, and the nature of task or the information-bearing object that contains the used or reused item.

Digital content and digital objects offer the possibilities to be used and reused in many different ways, contexts, and purposes. The Digital Library Federation Assessment Interest Group conducted a study on the use and usability assessment of digital libraries, emphasizing the importance of how users reuse digital content in various scholarly and non-scholarly contexts (Chapman et al., 2015). Their study found that reuse is associated with both the user’s intention and their behavior as they search for and identify information. It is apparent from the literature that evaluating reuse is heavily interconnected with a variety of points in the lifecycle of digital objects:
- discoverability of materials
- usability of both the digital objects found and the digital repositories in which they reside
- user’s motivation for seeking cultural heritage resources
- what users end up doing with the digital content

Chapman et al. also found that there is a growing number of studies that focus on how various disciplines, in particular humanities, reuse digital content such as images. In a study of Reuse of Wikimedia Commons Cultural Heritage Images on the Wider Web, Kelly (2019) notes that the reuse of digital cultural heritage media on social media platforms has received increasing attention in the scholarly literature over the course of the last decade.

In a study to assess users and and reuses of images from the Library of Congress collections, Reilly and Thompson (2017) concluded that ”... everyday users are repurposing digital content in ways that are meaningful to them, and they are acknowledging and fulfilling personal interests. These users are also sharing this content through a variety of environments on the Web, including popular social media platforms, blogs, and personal Web sites”. Other studies have investigated the intended use for digital objects in various libraries (Reilly and Thompson, 2014).

In 2017, a large scale survey of more than 300 cultural heritage knowledge organizations was conducted to develop a framework for the reuse of digital objects. The study found that 40 percent of question respondents indicated that they collected reuse statistics. The most common method for collecting reuse data was “Social media metrics”, “Alert Services”, and “Reverse image lookup”. The most common kinds of reuse data collected were “Digital collections and objects cited in scholarship” and “Published or re-posted digital objects in digital media”. The least common kinds of reuse data collected were “Published data sets used in new research” (Kelly et al., 2018).

3.0 Methodology

In this study we made use of a combination of methods to empirically develop a faceted conceptualization of digital object reuse. We used the results of a survey of digital repository managers on digital object reuse, which constitutes part of a research
3.1 Survey Method

The follow-up survey asked digital repository practitioners to prioritize a set of use cases aimed at assessing digital object reuse. The use cases corresponded to three categories: data collection, analysis, and reporting; programs and services; and privacy, rights management, and ethics. The survey respondents identified five use cases as the highest priority (see “Setting a Foundation” for a specific list of primary and secondary use cases):

1. Understand how content is being reused in a variety of contexts by various audiences (social media, classrooms, scholarly works, genealogy, digital humanities, etc.)
2. Tell stories of impact with the reuse data that has been captured and tailor it to specific audiences/stakeholders
3. Assess quality and quantity of items reused to inform digitization projects and priorities
4. Enable/encourage attribution of materials in various reuse contexts, including through sharing and reposting on social media, integration into classroom instruction, citing in scholarly works, or through non-academic avenues
5. Know and understand digital repository users without violating user privacy

The project team draws upon these prioritized use cases as the foundation for building the assessment reuse toolkit (Kelly et. al., 2018; Thompson, et. al., 2019). But, the toolkit will draw upon all of the 9 use cases identified.

3.2 Facet analysis

Originally introduced by Ranganathan (1967), facet analysis is primarily a logical approach to classification and knowledge organization (Hjørland, 2013). Broughton advocated for the use of faceted structures as digital library management tools (Broughton, 2001) and previous research has made use of facet analysis for the development of Basic Concepts Classification (Szostak, 2017). In a discussion of the cultural dimensions of facets in knowledge organization systems, Smiraglia (2017) noted that facets provide impetus to improved information architecture and knowledge organization and make the world a better place by making classification more efficacious. Facet analysis has previously been used to provide a basis for the understanding and communicating the nuanced aspects of big data (Shiri, 2014). In this paper, we argue that the facet analysis technique can be used as a basis for the development of a comprehensive and logical conceptualization of digital object reuse and the categorization of its various types, contexts, and cases in order to facilitate shared understanding and communication.

4.0 Proposed Faceted Categorization of Digital Object Reuse

Given the complex nature of digital object use and reuse and the nuances associated with the ways in which decisions can be made between use and reuse, we are proposing a matrix that captures the levels of use, considering the extent of passive versus active interaction with digital objects. As can be seen from Table 1, three concepts of Use,
Reuse and Transformation are introduced to capture, conceptualize and clarify the notion of reuse in relation to use and transformation of digital objects.

Table 1. Concepts of Use, Reuse, and Transformation

<table>
<thead>
<tr>
<th>Use</th>
<th>Reuse</th>
<th>Transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passive interaction with a digital object that may indicate potential interest and/or low value to an external user</td>
<td>Active interaction with a digital object(s) that results in sharing, transmitting, or forwarding access to the object, thereby indicating potential significance or value to an external user</td>
<td>Active interaction with a digital object that alters or changes the digital object, resulting in a new, distinct entity and indicating high impact or value to the user</td>
</tr>
</tbody>
</table>

- Browsing digital repositories for content
- Clicking a link for a digital object
- Downloading digital objects
- Accessing a web archive
- Watch a video online
- Read an article
- Viewing a photograph
- Listening to a song
- Taking a picture with a camera phone
- Printing digital objects
- Adjust lighting or coloring of digital items in order to faithfully represent the original object
- Charting a data set in a graph or infographic to communicate with others
- Recording a book to make an audio book
- Annotation of an image or document
- Translating the text of a digital object from one language to another
- Transcribe a digital object
- Create closed captioning for a video
- Curated sets of digital material, such as People of Color in Medieval European Art History https://medievalpoc.tumblr.com/
- K-12 education kits
- Painting, drawing, or otherwise artistically representing a digital object
- Adding color to a black and white photos or video in order to add artistic value to the original object
- New editions of a textbook
- Updated application profiles
- Updated procedures or documentation
- Reproducing a research study using the original parameters of an existing study
- Creating then and now photographs for an exhibit
- Combining two or more data sets for analysis
- Creation of a GIF or meme from digital objects
- Overlaying a map with data points

Based on the conceptualization of the notion of reuse in the above table, we are proposing a faceted conceptualization of digital object reuse that takes into account the various nuances of the definitions and dimensions of digital object reuse. Table 2 provides an overview of the key facets, sub-facets and values for digital object reuse understanding. There are five key facets include **Content**, **Process** (transformation), **Goals/intention**, **Agent** (organizations and individuals), and **Environment**. These facets are based on the conceptual foundation of faceted classification. Each facet has its own sub-facets and values.
Table 2. A faceted classification of digital object reuse

<table>
<thead>
<tr>
<th>Facet</th>
<th>Sub-facets</th>
<th>values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content</strong></td>
<td>-Content type</td>
<td>-Image</td>
</tr>
<tr>
<td></td>
<td>-Document type</td>
<td>-Text</td>
</tr>
<tr>
<td></td>
<td>-Media type</td>
<td>-Audio</td>
</tr>
<tr>
<td></td>
<td>-Etc.</td>
<td>-Video</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Software</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Dataset</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Metadata</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Research data</td>
</tr>
<tr>
<td><strong>Process (transformation)</strong></td>
<td>-Automatic</td>
<td>-Aggregated</td>
</tr>
<tr>
<td></td>
<td>-Manual</td>
<td>-Analyzed</td>
</tr>
<tr>
<td></td>
<td>-Semi-automatic</td>
<td>-Curated</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Changed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Derived</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Enhanced</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Mash-ups</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Major transformation/ alteration</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Repackaged</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Reformatted (file or format, e.g. compression)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Synthesized</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Subtracted</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Unprocessed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Versioned</td>
</tr>
<tr>
<td><strong>Goals/intention (Use/Reuse)</strong></td>
<td>-Access</td>
<td>-Gaming</td>
</tr>
<tr>
<td></td>
<td>-Consumption</td>
<td>-Reading</td>
</tr>
<tr>
<td></td>
<td>-Sharing</td>
<td>-Reviewing</td>
</tr>
<tr>
<td></td>
<td>-Data capture</td>
<td>-Researching</td>
</tr>
<tr>
<td></td>
<td>-Data curation</td>
<td>-Shopping</td>
</tr>
<tr>
<td></td>
<td>-Data archiving</td>
<td>-Studying</td>
</tr>
<tr>
<td></td>
<td>-Data management</td>
<td>-Using</td>
</tr>
<tr>
<td></td>
<td>-Data preservation</td>
<td>-Viewing</td>
</tr>
<tr>
<td></td>
<td>-Data access</td>
<td>-Citing</td>
</tr>
<tr>
<td></td>
<td>-Data interoperability</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-Data discovery</td>
<td></td>
</tr>
<tr>
<td><strong>Agent (organizations, individuals)</strong></td>
<td>Software, algorithms, People</td>
<td>-Scholars</td>
</tr>
<tr>
<td></td>
<td>People</td>
<td>-Scientists</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Social scientists</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Humanities scholars</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Archivists</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Curators</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Librarians</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Records managers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Information managers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Researchers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Scholars</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Students</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Readers</td>
</tr>
</tbody>
</table>
| Shoppers          | -Database managers  
|                  | -Knowledge engineers 
|                  | -Information scientists 
|                  | -Data engineers 
|                  | -Data scientists 

| Environment       | -Recommendation systems  
|                  | -Social networks 
|                  | -Social media 
|                  | -Search engines 
|                  | -E-business sites 
|                  | -Data archives 
|                  | Institutional repositories 
|                  | -Digital libraries 
|                  | -Virtual organizations 
|                  | -Cloud-based systems and services 
|                  | -Mobile computing providers 
|                  | -Information providers 
|                  | -Harvesters 
|                  | -Data commons 
|                  | -Data centres 
|                  | -Collaboratories 
|                  | -Digital humanity projects 
|                  | -Classroom 

We are providing an example to elaborate on the relevance and usefulness of this type of conceptualization. Imagine a large academic library aims to measure the value, impact, and relevance of their digital collections based on not only digital object use, but also reuse of various digital objects in different contexts to inform collection development and community engagement strategies, and to introduce the ways in which the reuse of digital objects could be facilitated and tracked. In order to understand and identify various instances, examples, and manifestations of reuse of digital objects, the library has to establish what is digital object use, reuse and transformation, what are the contexts in which reuse can take place (for instance, classroom, art history, social media etc.), what types of interactions users get involved in when they reuse a digital object (for instance, simple browsing, manipulating data and, image, or transforming and recreating a new entity). The proposed facted conceptualization facilitates the understanding and communication of the instances, contexts, examples, and environments in which digital object reuse can take place. Future research needs to expand and enhance this typology to cover the more subtle and nuanced aspects of digital object reuse, its use cases and contexts. As part of the Digital Content Reuse Assessment Framework Toolkit (D-CRAFT) project, funded by the IMLS\(^1\), we are currently investigating the development of a toolkit that would support and enable

---

cultural heritage and knowledge organizations to conduct digital content reuse assessment by developing various use cases and tools.

5.0 Conclusion

The value and impact of digital objects and their reuse should be conceptualized not only as part of the scholarly communication lifecycle but also as part of lifelong learning, recreational experiences, and activities of digital information users and searchers. It is timely for the digital library community to ensure and promote the relevance and usefulness of digital libraries by providing new frameworks and measures that evaluate and assess the impact of use and reuse of digital objects in relation to intellectual and artistic creativity as well as to informed citizenry, social responsibility, and democracy.

Acknowledgement

This work is supported by the Institute of Museum and Library Services (IMLS) (LG-36-19-0036-19). The views, findings, conclusions or recommendations expressed in this article do not necessarily represent those of the Institute of Museum and Library Services.

References


https://doi.org/10.1080/19322909.2016.1223573.


https://doi.org/10.17605/OSF.IO/4YHFJ.


Carlos Guardado da Silva – Universidade de Lisboa, Faculdade de Letras, Centro de Estudos Clássicos, Portugal
Luís Corujo – Universidade de Lisboa, Faculdade de Letras, Centro de Estudos Clássicos, Portugal
Jorge Revez – Universidade de Lisboa, Faculdade de Letras, Centro de Estudos Clássicos, Portugal

The Classification Plan for Local Administration
Portuguese Archives and the Knowledge Organization in Practice

Abstract:
This paper presents recent developments in Portuguese local archives’ classification policies. Following several professional workgroups and reinforced with a political mandate, local administration archives are putting in practice a new knowledge organization approach through the development of a national Classification Plan for Local Administration (CPLA). New archival legislation is ready to be officially published and its impact is expected to largely modify administration practices. The main objective of the study is to discuss the knowledge organization solutions embedded in the creation of CPLA. To achieve this goal, a literature review was performed, together with a project management approach, exploring a qualitative analysis of the problems and the solutions found during CPLA’s design. The Functional Macrostructure for Public Administration information is the main framework for the development of CPLA, and contributes to the emergence of a new paradigm in Information Management. Such paradigm intersects functions with business processes, transversally and supra-institutionally, and considers the different public administration bodies as open systems. As an outcome, the public administration and, in particular, the archival community receive a new standardized information management tool that serves the classification, appraisal and selection of archival information.

1.0 Introduction
This paper presents recent developments in Portuguese local archives’ classification policies. Following several professional workgroups and reinforced with a political mandate, local administration archives are putting in practice a new knowledge organization approach through the development of a national Classification Plan for Local Administration (CPLA). New archival legislation is ready to be officially published and its impact is expected to largely modify administration practices.

The main objective of the study is to discuss the knowledge organization (Hjørland 2016) solutions embedded in the creation of CPLA. To achieve this goal, a literature review was performed, together with a project management methodology (Abadal 2004), exploring a qualitative approach of the problems and the solutions found during CPLA’s design.

The CPLA initiative has been regularly mentioned in Portuguese literature. There are papers written by authors from the archival policy coordinating body (Lourenço et al. 2019), written by members of the working groups or who have already performed locally early adoption of the new methodology (Neves 2014; Melo 2016; Negrão 2017; Torres 2018) and presented by scholars (Silva 2013, 2016, 2019).

The previous legislation regarding public information management had the advantages of streamlining procedures and costs through the elimination of worthless records that had fulfilled their administrative function. However, the old legislation tended to be applied downstream of its production, repeatedly causing an entropy in the public administration services. In the new scenario, the adoption of more objective criteria and
a relational methodology established between business processes strongly recommend that appraisal should occur in a genetic phase, thus enhancing the continuous management of information flows that result from the internal and external procedures, from the moment of their production until their permanent conservation or permanent disposal (Silva 2016).

2.0 Functional Macrostructure

In an attempt to create a set of tools of information classification for its entire public administration, a Portuguese collective initiative conceived a new structure with several theoretical and practical implications, called Functional Macrostructure. The process was led by DGLAB (Directorate-General of Book, Archives and Libraries), the coordinating body for the national archival policy, and involved more than two hundred public administration organizations (central, regional and local) since 2004 (Silva 2016).

The Functional Macrostructure is a conceptual representation of functions performed by public sector organizations, which develops on two levels: (a) Top-level instances representing public administration functions; (b) Second-level instances representing the subfunctions into which the first can be decomposed. It has a global and integrative perspective of the public sector, which overlaps the analysis of the functions of each single organization, taking as greater value the semantic interoperability in the informational exchanges of the administration bodies. The Functional Macrostructure aims to increase semantic interoperability in information exchange between public administration services, responding to the need for a common language for information representation, and ensuring the ability of a system (digital or not) to communicate transparently with another system. It also aims to provide a classification model (in syntax and semantics) that can be understood by different entities of the public administration. Finally, another goal is to promote the use of a single structure to classify and organize public sector records (and archival information that does not materialize in documents, and which it is also necessary to manage), enhancing greater management effectiveness and easier access to information by bodies and citizens, and contribute to transparency in management and to the enhancement of accountability processes.

Since 2012, the Functional Macrostructure achieved a new development stage, surpassing the central Public Administration scope. The goal was to reach local administration governing bodies. This evolution was mainly due to the initiative of a set of municipalities that organized themselves into a cooperative project to carry out a feasibility analysis of the application of the Functional Macrostructure to the functional framework of municipal services and, in the face of a predominantly positive response, to develop a classification plan for local administration.

This new stage underlined the original purpose of the Functional Macrostructure, designed primarily to support functional classification plans. In 2017, DGLAB published the first version of the Consolidated List for classification and appraisal of public sector information. This list included already a set of third-level instances representing the business processes and it was the result of several parallel projects concerning the Functional Macrostructure development process. The document presented a hierarchical structure of classes that represent the functions, sub-functions and business processes performed by the Public Administration, from a supra-institutional, cross-functional and
functional perspective, including their description and appraisal (administrative retention periods and the final destination of information). It was proposed as a reference for the development of organizational or multi-organizational tools for the classification and appraisal of public information (classification plans and retention schedules).

Since 2018, the Consolidated List was published as open data through the CLAV (Classification and Appraisal of Public Information) Platform. This new tool is ontology-based, including the business processes of entities that perform public functions (Consolidated List), associated with a catalog of the legislation that regulates them and the bodies that execute them, that are responsible for the preservation of the information produced in this context and for the lifecycle management, and retention schedules, derived from the Consolidated List, identifying the classes and information lifecycle associated with business processes (administrative retention deadlines and final destination) for implementation in organizational or multi-organizational context (Lourenço et al. 2019).

3.0 Classification Plan for Local Administration

Referring to the functional dimension, the classification consists of identifying and grouping records and information into classes or categories according to the functional structure and their sub-functions or activities of the various public administration producers, regardless of their nature. However, the adoption of the functional classification is complex because it is based on the clear distinction between bodies, divisions or services and functions, when administrative units are established, or should be established, based on functional specialization. However, function-based classification is clearly the most stable criterion, being as objective as the organic one if functions are clearly determined (Silva 2019).

The new CPLA, made up of a functional matrix, with the version 0.2 already published in 2014 (Direção-Geral do Livro, dos Arquivos e das Bibliotecas (DGLAB) 2014), is together with the Functional Macrostructure (version 2.0) a decisive management tool for the emergence of a new paradigm in local government archival information management, as it is connected with the universe of all public administration. It was considered advantageous to implement more extensive solutions for the organization of the Administration's information and its interoperability. Instead of each entity creating its own classification plan, the use of an instrument by all entities, safeguarding the specificities of each one, not only guarantees an economy of scale, making public investment more profitable, but also facilitates easier access to information by organizations and citizens (Lourenço and Penteado 2015).

In response to the continuous challenges of administrative modernization and e-government, new policies and instruments for the management of records are required, a requirement that also extends to classification plans. These, for a long time, seem to not respond effectively to the new challenges, if they continue to be considered only as a tool for organizing and describing documentary fonds. Even today, many municipalities do not use any classification plan in the management of daily records. On the contrary, the classification plan should be a key instrument in the management of records, as well as reflecting the functions and activities in each organization. If its importance was recognized in traditional information systems, when they existed, in digital information systems, classification is a mandatory and indispensable requirement (Silva 2019).
The CPLA was developed according to the Functional Macrostructure. The first premise for the elaboration of a classification scheme adapted to the reality of the municipal archives was the recognition that there are transversal processes to several public administration bodies. In this sense, the CPLA Working Group, formed in September 2011, by the municipalities of Cascais, Lisbon, Lourinhã, Oeiras, Santarém, Sintra and Torres Vedras, as well as the Municipalized Services of Oeiras-Amadora, with the technical support of DGLAB, and a year later, by the municipalities of Amadora, Alenquer and Beja, concluded that the functions of the central administration are essentially identical to those of the local administration.

Following this starting point, the Working Group proceeded to the identification of specific (from Local Administration), common (to Central Administration) and transversal (between different Local Administration bodies and between Local Administration and Central Administration) business processes. At the end of the first phase (Sept. 2011-Sept. 2012), the identification of business processes, version 0.1 of the Classification Plan was published, putting the document up for discussion to collect contributions from the various bodies of local administration.

In the second phase (Sept. 2012-Feb. 2013), the group was expanded with more 45 municipalities as observing members, with online access to the files created so that, in an articulated way, they could send their contributions. Then, a third phase (Feb. - May. 2013) started, through the creation of groups by function, bringing together professionals from local administration and central administration, aiming at the consolidation of common and transversal business processes, in order to standardize their identification and to facilitate interoperability. The fourth phase was the consolidation of the Classification Plan for the public release of version 0.2. The fifth phase will follow, which aims at associating the conservation periods and the final destination with the business processes, and which will introduce the fourth level in the classification code, for the purposes of information appraisal and selection (Silva 2019).

During this course, the project's initial assumptions were maintained, namely: (a) Creation of a single tool with concerns in terms of classification and appraisal; (b) Construction of an instrument that promotes semantic transversality and interoperability; (c) Application of the classification plan to the records (and information) at the stage of its creation or reception. It should be stressed that this plan is not an instrument of retroactive application, nor does it intend to reorganize the documentation already produced; (d) Use of the instrument for the classification of records on paper or other media. It applies to electronic records management systems, as well as electronic information systems in support or business areas; (e) The implementation of the classification plan necessarily implies the adoption of new practices, which may lead to a rearrangement of records organization (Direção-Geral do Livro, dos Arquivos e das Bibliotecas (DGLAB) 2014).

4.0 Knowledge Organization (KO) solutions

Classification is a term concerning «the assignment of elements to pre-established classes. The wide meaning of classification is the process of distinguishing and distribution kinds of "things" into different groups» (Hjørland 2017). Grounded in theory, CPLA follows a process approach, traditionally more claimed than practiced, but con-
considering the business process concept. This has required the standardization of terminology, since it is a recent concept in the context of information management, as its meaning is not univocal, nor is it entirely acknowledged by archivists (Sabourin 2001; Díaz Rodríguez, Caro Castro, and Hernández Oliveira 2010).

A hierarchical semantic classification structure of the CPLA was developed at three levels, the first two representing the functions and sub-functions of public administration (Functional Macrostructure) and the third level representing the Business Processes. Following KO theory, classification is the intellectual operation for the identification and hierarchical structuring of documents and information of an open archival system that reflects the functions and activities of the agencies involved in the different business processes, based on the principles of interoperability, completeness, transversality and suprainstitutional (exosystemic) approach (Foscarini 2009; Barros 2012; Simões and Freitas 2013).

The fonds extends to the transversality of organizations’ processes in suprainstitutional terms. Neither the principle of provenance seems to fit here, since it refers to the uniqueness of the fonds, given to a single producer. As a new paradigm in the archival information management of the public administration is initiated, which considers, in the transversal processes, two or more producers, the concept of background was closely related to the competences of one producer, despite the contradictions between these two concepts.

The CPLA, being a plan grounded in Functional Macrostructure, integrates the classes of 1st and 2nd level (representation of the functions and sub-functions of the public administration) and develops the 3rd level also in a functional perspective. The 3rd levels represent, as much as possible, the business processes developed by local administration. Eventually, due to needs derived from the appraisal, 4th levels may be created.

The identification of business processes implied the adoption of the following principles: (i) Interoperability (existence of communication between systems); (ii) Transversality (existence of reciprocal relationships); (iii) Granularity (subdividing a system into smaller parts); (iv) Completeness (complete presentation); (v) Exogeneity (existence of supra-institutional or exosystemic value chains); (vi) Respect for the function (hierarchical dependence on the business process of the 1st and 2nd level classes) (Silva 2019).

In order to obtain a single classification plan for the entire Portuguese public administration, regardless of different materializations, it was necessary to represent the logical path followed in each function for obtaining business processes. Thus, the option was the construction of “concept maps”, which would allow to guide, explain and justify the path. This information does not belong to any of the classes, as it is not directly visible in the Classification Plan.

For the subdivision, in order to structure the classification system, the principles or types of semantic relationships between the concepts, proposed by Dahlberg (1978), were followed, namely:

• Hierarchical relationship or gender and species relationship (when two or more concepts in the subdivision have identical characteristics, and one of them has one more characteristic than the other, that one is more specific, the other is broader);
• Partitive relationship (when between the whole and the parts or between a product and the elements that constitute it);
• Opposition relationship (when contradiction or contrariety);
• Functional relationship (when a function derives from a process, based on the semantic valence of verbs).

4.1 First KO solution – Local Administration working group and Coordinating Body leadership and support

Following the initiative of a set of municipalities and their archivists, the first KO solution was the informal working group gathered to work on CPLA (Melo 2016). Although it seems a managerial option, this initiative exemplifies the social dimension of information systems and networks. Without the involvement of local archives’ professional group, it would be quite difficult to connect the work performed in the central hub with the local nodes. The CPLA design was possible based on concerted will, initiative and “appetite” to build a tool that fits the new information management models in those organizations (Melo 2016).

Neves (2014) also highlighted the collaborative work and the harmonization effort between Central and Local Administration entities. This methodology is considered both a condition and a guarantor of the greatest value of the tool: its supra-institutional and supra-organic characteristics. The ability of the different Public Administration entities to interact, sharing information within the scope of the processes in which they intervene, requires the existence of tools for interoperable solutions. However, tools that allow interoperability are not built without articulation between entities that share the same concerns, that intervene in the same processes, that compete together for a service or a final product (Neves 2014).

The aim of the project was never just to build a classification plan for the local administration: from the beginning, it was assumed that the plan would comply with Functional Macrostructure, which implied adopting and assuming its principles and concepts. It therefore means that the objective of the project is based on the promotion of semantic interoperability that serves the objectives of entities that share responsibilities in the provision of public services. The principle of supra-institutional Functional Macrostructure required to abandon the perspective of Local Administration and to seek the perspective of Public Administration (Neves 2014).

As with any project, a working methodology was determined, which, at first, was exposed by DGLAB. As the work evolved, that is, when the theory was confronted with the practical reality, some concerted revisions were made between DGLAB and the working group. However, strictly speaking, the main guidelines for the survey / classification of Business Processes resulted from the reflections and the work of several years developed when the Functional Macrostructure was built (Melo 2016).

4.2 Second KO solution – Functional-based classification

Classifying is the intellectual act or operation that consists of grouping elements, which have something (characteristic or property) in common, forming a set or a class. The act of classifying allows the creation of classes of elements with connections to each other and distinguish them from other classes that do not contain the same characteristics (Ribeiro 2013). Classification is also the act of identifying and organizing archival information by class and / or category according to the functional structure and respective
The implementation of the CPLA supposes the adoption of a functional approach, by business processes, which may imply changes in administrative practices and, consequently, in the organization of information (Neves 2014). The appraisal process focuses on the attribution of value and on the determination of the producing entity responsible for the retention of the records in the immediate moment, regardless of the reuse of the information (it is appraised in response to what it is; information is valued due to its context of production and not to the potential uses of that information) (Lourenço and Penteado 2015).

The weight of organizational culture, based on management models that have privileged its umbilical or self-centered view, is contrary to the principles on which Functional Macrostructure is based and which are characterized by the supra-institutional or supraorganic approach. This interorganizational approach allows entities to integrate and share services and information in the pursuit of cross-cutting activities, aiming at the effectiveness and efficiency of Business Processes (Melo 2016).

Only the functional approach allows for the necessary distance from the vices rooted in the tradition of a hierarchical and overly departmentalized administrative model, favorable to the fragmentation of processes, the redundancy of information and duplication of tasks, being therefore adverse to the sharing of information and competitiveness of companies, public organizations. One of the first consequences resulting from the adoption of the functional approach was the redefinition of the methodology used to identify third-level classes. For the purpose of identifying third-level classes in a classification plan built on a functional logic, it is necessary to capture something more stable and common to the different entities, it is necessary to know what they do and not how they do it. It then became evident the need to identify not the documentary processes but the business processes that result from the performance of public functions by the different entities (Neves 2014).

The synergies created in the field of public information management seeks to break with the traditional and obsolete organizational information management, a model that does not comply with the requirements of a tendency to digital administration, capable of responding within acceptable deadlines. Definitely, information management, that is, its classification and appraisal, involves models and tools that are integrated and shared by public agents, at the suprainstitutional level (Melo 2016).

4.3 Third KO solution – Business process approach

The appraisal of information resulting from business processes will imply the adoption of certain principles, among which: (a) Functional provenance (the information produced is no longer linked to a particular producer, being the result of the performance of several producers who work in networks - open systems -, relating more to the functions and activities and less to the organizational structure). Such principle will allow explaining the origin and evolution of organisms, as well as the way they act or interpret social reality, focusing on the functional context; (b) Globality (transversal approach to public administration, with the information produced being appraised in its entirety, regardless of the medium); (c) Legality (appraisal of information resulting from the application of the legal regulation); (d) Objectivity and contemporaneity of the judgment
(appraisal according to principles, accepted social values, knowledge and technology available at the time of records’ creation); (e) Participation (appraisal of information based on the definition of the owner of the business process, as well as on the existence of customer relationships, when existing or other forms of participation); (f) Responsibility (appraisal and selection actions are based on defined and known criteria, with disposal events being demonstrable); (g) Exceptionality (selection of information resulting from exceptional events such as natural disasters); (h) Informational density (nature of information, requiring retention of synthesized information); (i) Partition (when different disposition dates and / or disposal decisions are defined for the information, depending on the stage in the business process); (j) Authenticity / integrity (appraisal of authentic documents); (k) Guarantee (safeguarding the rights of the producing entities and content object) (Silva 2019).

Each member of the working group had to carry out an in-depth study of the underlying principles and values, so that the survey of business processes corresponded to the “spirit” that guided the construction of these tools that promote semantic interoperability. After this analysis, the 19 Functional Macrostructure functions were then distributed among the participating municipalities (Melo 2016).

The identification of business processes for the constitution of classes is something that stems from the functional approach and requires the adoption of different work methodologies. The Business Process must be identified within the scope of the function to which it belongs (what it is and not for what it serves or how it is), it must be identified taking into account the final product and not the partial or intermediate product, it must not consider the specific procedure for obtaining that product (given that the procedure may vary from entity to entity, or due to legal imperatives) and must take into account the specific legal framework, although it is recognized that there is no univocal relationship between a Business Process and the legislation. The concern in associating the final product with the concept of Business Process led to another principle underlying this project: the principle of transversality that guarantees the identification of the Business Process in its completeness. This is only possible with a supraorganic and supra-institutional approach (Neves 2014).

The work methodology for the survey / classification of business processes is generally characterized in three stages: (1) Investigation work by municipality; (2) Meetings with producer services; (3) Meetings between municipalities and DGLAB. Subsequently, DGLAB proposed the mediation of a supra-institutional harmonization phase of business processes, which took place at a meeting between central and local government entities (Melo 2016).

The construction of the CPLA allowed a structured view of public administration, from the perspective of a single big system, in the articulation of multiple public systems, which tend to be open, shareable and flexible, to the detriment of the traditional positioning in the closed, isolated and inflexible organizational system. Based on the focus on supra-institutional relations established between business processes, which represent a network of highly complex transactions, forming a web in which the public administration itself is a recurrent victim, it seeks to clarify the different actions, through the creation of a semantic language that enhances the understanding between services and public bodies in the search for the satisfaction of external and internal customers (Melo 2016).
5.0 Conclusion

This work showed how a common, shared and standardized language was created to represent the functions of public administration, called Functional Macrostructure. In order to deepen the levels of semantic interoperability achieved in this structure, DGLAB developed a Consolidated List for the management of decisions on the classification and appraisal of public information.

A single tool for classifying and appraising archival information is available to all entities of the Portuguese local administration and also other public administration entities. Nowadays, it seems increasingly difficult to understand the classification as a function that deals with the documents of a fonds, this is produced within the scope of an organization's activities, based on the archival principles of provenance and respect for the original order. The standardisation of processes will have to be definitively inserted in the production phase, contributing to an effective management of information and records in the information systems of public administration bodies (Silva 2019).

The paradigm proposed by Portugal is characterized by a new archival approach, a matrix model that crosses functions with business processes. The integration of traditional documentary series into the Functional Macrostructure is irreconcilable. Such change definitely positions the information manager in the design, planning and development of the information system. Finally, the Functional Macrostructure, as well as the CPLA, contribute to the emergence of a new paradigm in Information Management within the public administration. Such paradigm intersects functions with business processes, transversally and supra-institutionally, and considers the different public administration bodies as open systems. As an outcome, the public administration and, in particular, the archival community receive a new standardized information management tool that serves the classification, but also the appraisal and selection of archival information.

It is expected the publication of the legal diploma that will regulate the classification, appraisal, selection, elimination and conservation of information, in any medium, in the exercise of local administration functions. The administrative procedures associated with them are included, thus streamlining the functions of the archive in guaranteeing rights and duties and in preserving collective memory.

References


Identifying and Classifying the Phenomena of Music

Abstract:
The classification of music for information retrieval has a long history, predominantly associated with the distribution of printed music in classes based on musical medium and form. Recent research has delved into specific aspects of the classification of music such as performance and reception, in addition to the finer aspects of medium and form. Meanwhile, new input from the music information retrieval community has pointed to the potential richness of music classification that takes into account a range from simple aboutness to more auditory concepts such as listener emotion, holistic user experience, or task complexity. The extension of the classification of music in the Basic Concepts Classification requires a larger embrace of musical phenomena. A large array of musical phenomena is identified, leading to a flexible but exhaustive system of facets, and documenting the grammar of a facet analytical approach to classification of musical phenomena. A synthetic approach within a general (universal) classification can facilitate classification along diverse dimensions such as the subject of a work, the composer’s intentions, and the intended audience.

1.0 Introduction: Classifying music
The classification of music for information retrieval has a long history (Smiraglia 1989; McKnight 2002). Much of the richness of the history of the creation of music classification schemes arises from the professionalization of music librarianship in the United States and United Kingdom from the early twentieth-century forward. By the mid-1950s the growth of specific practices in music libraries was synthesized as the distribution of printed music in classes based on musical medium and form (Meyer-Baer 1951 [1973]; Elmer 1957 [1973]). Meyer-Baer contrasted the broad categories (church music, vocal music, keyboard music, etc.) of the Dewey Decimal Classification with the granular medium-based arrays of the Library of Congress Classification: M, and then placed those over and against a simple pragmatic local classification that allowed the addition of style period indicators. A hallmark of music classification was the 1938 Dickinson Classification of Musical Compositions, originally developed at Vassar College but eventually used also at Columbia University and The Juilliard School. Dickinson’s classification is medium-based, but uses a complex system of composer “book numbers” to create alpha-numeric arrays of a composers’ works within a class, and also somewhat precociously makes use of what we now call facet analytical theory by permitting the addition of symbols and letters to introduce faceted indicators for arrangement, voice range, excerpt, etc. Sound recording collections, especially those in public libraries, also contributed what now might be called “best practices” by generating highly pragmatic classifications that mimicked those of record stores, in which bins based on broad themes--e.g., Operas, Piano, Musical Shows, Jazz Music, Holiday Music, etc.--allowed library users to browse through and select among LP recordings (see “ANSCR” in Smiraglia 1989, 114ff.).

Recent research has delved into specific aspects of the classification of music such as musical performance and reception (Lee 2011, 2015), in addition to the finer aspects of medium and form, including musical medium and music ensembles (Lee 2017a, 2017b; Lee and Robinson 2017). The idea of a performance as an entity separate from
the musical work, its printed instantiations, or recordings of its expressions, is critical
but has received only little attention. Smiraglia (2007) demonstrated empirically the
instantiation network of a set of performances, which can be thought of as “works”
distinct from the musical abstractions in them. Twelve years after this paper only a few
scholars have thought to distance themselves from the error of considering a
performance to be a direct instantiation of a work. Cruz and Smiraglia (2020), who work
with Brazilian popular music, is a notable example. For them performance of a “musical
idea” instantiated through both “arrangements” and “performance expressions” is fully
modeled without reference to what would be subsequently-created notated documents
or recordings.

Attempts to generate more flexible systems of facets for musical concepts and rules
for their combination also point to potential richness of the classification of music
phenomena. The complete revision of music schedules in DDC in the 1980s was
undertaken with a facet analytical theory in mind (Sweeney 1990). The complete set of
facets arrived at included: theory, elements, techniques, character, forms, executant,
composer. The use of the base DDC music schedules for both notated music and books
about music was accomplished by designing two different citation orders; the order for
music itself was “executant, forms, character,” with the expectation (mirroring
Dickenson) of the subsequent addition of a composer-facet symbol to create
alphabetico-synthetic arrays of works under specific executants. One distinction that
arose in implementation was to shift the citation order for vocal music to “forms,
executants, character.” A thorough analysis of the rules for generating Library of
Congress Subject Headings (LCSH) for music was outlined by Young (1998). At the
time, the use of LCSH for music required catalogers to combine terms from simple lists
of medium and form with indications of number to create otherwise uncontrolled
headings. Based on the contents of the cataloged artifact (score, recording), the heading
could either have form or medium as lead term, to which subdivisions for the other were
added. Additional subdivisions for physical form, and occasionally period of
composition, were allowed. Young’s detailed instructions cover every aspect of what we
will later call “grammar” of music facets. A 2015 paper by Madalli, Balji and Sarangi
applied ontological analytical concepts to the domain of music to generate a set of facets
for a music ontology: these were “theory, person, instrument, kind, form, work.”
Iseminger et al. (2017, 430) describe the evolution of thesauri from elements of the
former LCSH, revealing potential thesauro-facet arrays for the usual suspects—topical
headings, genre, form [and] medium. Meanwhile, new input from the music information
retrieval community has pointed to the potential richness of music classification that
takes into account a range from simple aboutness to more auditory concepts such as
listener emotion—e.g., amazement, solemnity, tenderness, etc.—(Aljanaki, Wiering,
and Veltkamp 2016), holistic user experience (Hu et al. 2015; Downie 2003)—e.g.,
boring, indifferent, hopeful, circumstance, etc., or task complexity—e.g., lyrics,
translation, buy or download, etc. These IR approaches are particularly important for a
classification of music that might be used for semantic web (SW) applications.

2.0 Grammar for faceting

Szostak (2017a) has described an approach to faceting that uses simple grammar to
reported on the exploration of this wide variety of approaches to classifying music within the Basic Concepts Classification (BCC). Szostak (2019) explores the general advantages of a synthetic approach to classification, with particular attention to the classification of music.

Since the BCC has separate schedules of things (mostly nouns), relators (verbs and conjunctions), and properties (adjectives and adverbs), the subject headings formed in BCC tend to resemble sentence fragments. Though such subject headings may surprise those used to the ungrammatical format of most subject headings in the world, there are huge advantages to a grammatical approach. First, humans spend most of their lives thinking in sentences, and can thus more readily comprehend a subject heading that is expressed in grammatical format. Second, linguists appreciate that sentences clarify the meaning of terms within a sentence. A grammatical approach thus further decreases linguistic ambiguity (and BCC terminology is generally terminology that has broadly shared understandings across disciplines and groups).

Third, the nature of a work is the ideas it expresses (see Smiraglia 2001), and these are expressed in one or more sentences, often of the form X has effect N on Y. User queries are generally also expressed in a sentence. We can do a better job of guiding users to documents if we translate the user query into a sentence-like subject heading, and likewise translate the key idea of a work into a sentence-like subject heading. We at present go from a sentence-like query to an ungrammatical subject heading to a work best defined by a sentence.

Fourth, Szostak (2017b) showed how all of the facets identified within both the Bliss Classification and the Integrative Levels Classification can be interpreted as either distinct elements of a grammatical sentence, or as distinct schedules within the BCC classification of phenomena. The BCC thus clearly expresses all facets without needing to devote notational space to facet indicators. The classifier need not explicitly perform facet analysis, but can merely translate a sentence in a document description into BCC terminology. They can, if they wish, easily check to see which facets were addressed.

3.0 Methodology: The domain analysis clinic

The extension of the classification of music in the BCC is an essential part of the Digging into the Knowledge Graph research project,1 in which the classification of specific musical concepts rather than physical musical documents requires a larger embrace of musical phenomena. Here we describe specific work undertaken to define a larger array of musical phenomena, to generate a flexible but exhaustive system of facets, and to document the grammar of a facet analytical approach to classification of musical phenomena. In November 2019 a small group of experts in the classification of music was assembled at the Institute for Knowledge Organization and Structure, Inc. (IKOS) in Lake Oswego, Oregon (USA). The group constituted what IKOS has called a “domain analysis clinic” (DAC) on “the phenomena of music for classification.” The general outline of a DAC includes an invitation-only group, assigned “homework” to build an exhaustive corpus of relevant research from which segments of meta-analysis

---

1 Digging Into the Knowledge Graph.
are generated. When the group meets the meta-analysis is reviewed, synthesis is constructed, and follow-up assignments are fixed with the purpose of filling identified gaps in knowledge of the specific domain (Smiraglia 2019). Szostak and Smiraglia (2019) focused on how a synthetic approach within a general (universal) classification could facilitate classification along diverse dimensions such as the subject of a work, the composer’s intentions, and the intended audience. Participants in addition to Szostak and Smiraglia were Deborah Lee, Richard Griscom, J. Bradford Young and Joshua A. Henry. Specific details of the meta-analysis and the generation of facets are reported in Szostak et al. (forthcoming). What follows here is the general outline of the fleshing out of schedules of musical phenomena for the BCC.

4.0 Musical phenomena in faceted arrays

Upon review of the meta-analytical data, the group arrived at the following set of musical phenomena that should be developed or extended for the BCC:

Character, occasion and function of the music
Types, forms and genres of music
Medium of performance
Commercial elements of recorded music
Format (arrangement, transcription, transformation, etc.)

In addition, consideration was given to traditionally relevant concepts such as the personal names of creative contributors (composers and librettists, but also sound editors, producers of performances, etc.) and to representations of place and time. BCC already allows synthesis of names, places and time designations.

4.1 Form, genre, etc.

It was decided to combine the elements identified above as “character, occasion, function, type, form and genre” into a single facet. The structure of this facet is to be based on the Library of Congress thesaurus for form and genre terms (LC Genre/Form Terms or LCGFT). LCGFT is maintained as linked open data (LOD) by the Library of Congress, with ongoing input from the active library community, including the Music Library Association (Library of Congress 2020):

“The Library of Congress Genre/Form Terms for Library and Archival Materials (LCGFT) is a thesaurus that describes what a work is versus what it is about. For instance, the subject heading Horror films, with appropriate subdivisions, would be assigned to a book about horror films. A cataloger assigning headings to the movie The Texas Chainsaw Massacre would also use Horror films, but it would be a genre/form term since the movie is a horror film, not a movie about horror films. The thesaurus combines both genres and forms. Form is defined as a characteristic of works with a particular format and/or purpose. A “short” is a particular form, for example, as is “animation.” Genre refers to categories of works that are characterized by similar plots, themes, settings, situations, and characters. Examples of genres are westerns and thrillers. In the term Horror films “horror” is the genre and “films” is the form.”

Some of the genres identified by the Library of Congress would be treated differently by the BCC. Most obviously, “humorous” is not really a distinct genre but a property that might be attributed to music from many different genres. One of the beauties of the synthetic approach taken by BCC is that terms from non-musical schedules can be used as necessary in the subject headings for works of music. “Humorous” is already a property within schedule Q of the BCC.
LCGFT is not music specific but the group easily extracted musical phenomena from the list, which can form the basis of a hierarchical array for BCC. LCGFT does include terms relating to styles and kinds of music. A decision also was taken that occasions, functions, and character could be synthesized by adding terms from elsewhere in the BCC. The BCC already contains schedules that encompass various types of celebration, group, organization and time period. One schedule that we hope to expand upon is the schedule CR regarding religion. We intend to identify in more detail the various kinds or parts of religious services, e.g., baptism, offertory, etc.

4.2 Medium of performance

Traditionally the basis of most music classifications, medium of performance is obviously an essential facet. The BCC has imported the Hornbostel-Sachs instrument classification. This classification attempts global coverage, and provides a hierarchical structure grounded in the physical characteristics of instruments. However, the taxonomical terms used are not particularly directly the names of the “phenomena” of musical medium. For example, “flute” is embedded in a hierarchy of “aerophones,” and “piano” is under “pianoforte” embedded in a hierarchy of “chordophones.” The group urged incorporation of the Library of Congress Medium of Performance Thesaurus for Music (LCMPT), which like LCGFT is maintained in consultation with the Music Library Association (http://id.loc.gov/authorities/performanceMediums.html):

“The Library of Congress Medium of Performance Thesaurus (LCMPT) is a stand-alone vocabulary that provides terminology to describe the instruments, voices, etc., used in the performance of musical works …. Authorized terms and references in LCMPT generally consist of single words and phrases, but parenthetical qualifiers are occasionally employed to differentiate among homonyms. All terms and references are in the singular form ... (e.g., flute; saxophone ensemble; but Irish harp). The thesaurus has a few broadest terms as listed in the “Top Scheme Members” section. Each of the other terms is hierarchically subordinate to one or more of these terms and exhibits the class/class member relationship. Most of the authorized terms have Used For (UF) references for synonyms. Scope notes are also provided in many cases, and may describe the medium’s physical structure, the time period in which it was popular, and/or its geographic origin.”

For BCC, the group encouraged harmonization of the existing BCC schedule with LCMPT, and this task was assigned for work in early 2020.

4.3 Audiography

A new facet was outlined broadly with regard to input from the IR and SW communities. The general structure of the facet is to include:

- Details of capture (i.e., where and when was a performance recorded)
- Details of production and dissemination (release, music recording number, etc.)
- Physical or digital format (soundtrack, single, compilation, track number, etc.)
- User’s purpose: settle a bet, gift, etc.
- Emotion invoked by the music

User studies have shown that the entities on this list are those often sought by people looking for music online. Perhaps the most controversial part of the group’s discussion, this facet was tasked for detailed explanation in early 2020. It is worth noting that details of capture, emotions and purposes likely can be synthesized from existing arrays in the BCC.
5.0 Conclusion: Toward the grammar of faceted music classification

Classificationists can usefully ask what sort of queries a user might have. We might reasonably expect that users will want to search for works from a particular genre, or about a particular subject (love songs, say), or with a particular purpose (revolutionary songs), or for a particular occasion (wedding songs), or with a particular melody (maybe to accompany a particular video), or to invoke a particular emotion. And think of a group of musicians that want to play together and thus seek works designed for the particular set of instruments that they play. The simple fact is that all of these searches are either difficult or impossible within existing approaches to music classification. Szostak and Smiraglia (2019) detailed how the synthetic approach of BCC facilitated the classification of works by subject, many occasions, multiple creators (for when a work is rearranged), and many aspects of culture. The present project seeks to develop new schedules that will further enhance the classification of music. Though challenges remain, we are confident that we can satisfy user queries much better than is possible at present.

References


Interdisciplinary Searching as a Use Case for Vocabulary Mapping

Abstract:
There is increasing recognition of the importance of interdisciplinary research, but it is not well supported by available discipline-oriented information systems. The publication of ISO 25964-2: Information and documentation - thesauri and interoperability with other vocabularies - Part 2: Interoperability with other vocabularies (ISO 2013) brought into focus the principles and practice of vocabulary mapping, a possible approach to better support interdisciplinary searching. This paper reviews the challenges of interdisciplinary searching, the specifics of vocabulary mapping, and approaches to evaluating the resulting mappings. More studies assessing the retrieval performance and usability of mappings are needed in order to demonstrate the ways in which mappings can add value in the search process, especially for interdisciplinary searching.

1.0 Introduction
The ISKO 2020 conference theme of “Knowledge Organization at the Interface” relates to the challenge of facilitating searching at the interface among disciplines. As Dextre Clarke (2019) and Zeng (2019) make clear, the publication of ISO 25964-2: Information and Documentation - Thesauri and Interoperability with Other Vocabularies - Part 2: Interoperability with Other Vocabularies (ISO 2013) brought into focus the principles and practice of vocabulary mapping. But given the effort required to accomplish mapping among two or more vocabularies, it is important to ask, echoing Ford (2019) in his advocacy for use cases, “Who will use this and why?” This paper explores interdisciplinary searching as an important use case for vocabulary mapping.

There is increasing recognition of the importance of interdisciplinary research, as noted by Gibson (2012, 213): “The structures of the academy are straining toward integration across disciplines in order to solve transdisciplinary problems and ‘grand challenges’.” However, as Palmer and Fenlon (2017, 429) observe, “we are far from realizing the potential of information systems and services for fueling interdisciplinary research.” In order to better support interdisciplinary research, there is a need for continuing evolution of discovery tools.

But what type of research does the label “interdisciplinary” encompass? Wasserstrom (2006) notes a range of possibilities, including: team-based interdisciplinarity, to which several scholars bring different skills; cross-over interdisciplinarity, referring to fields like bioethics which have roots in two disciplines; and exploratory interdisciplinarity, for scholars who apply material from other fields on occasion. The term may also encompass explicitly interdisciplinary fields such as area studies, ethnic studies, gender and women’s studies, and environmental and resources studies. As explained by Linköping University’s Institutionen för Tema (Department of Thematic Studies), which encompasses Child Studies, Gender Studies, Environmental Change, and Technology and Social Change:

In an increasingly complicated world, the need for interdisciplinarity is greater than ever, researchers that can see both breadth and depth in major societal issues. At the Department of Thematic Studies, natural science, social science, technology and the humanities meet in the common aim to increase
In this context, interdisciplinary research attempts to integrate insights from multiple disciplines to address a problem that defies explanation by a single discipline. This has implications for education as well. A recently published report (National Academies of Sciences, Engineering, and Medicine 2018, x) observes that there is a concern that “an education focused on a single discipline will not best prepare graduates for the challenges and opportunities presented by work, life, and citizenship in the 21st century” and recommends “an approach to higher education that intentionally integrates knowledge in the arts, humanities, physical and life sciences, social sciences, engineering, technology, mathematics, and the biomedical disciplines”.

Because the most widely used knowledge organization systems were developed “when a discipline-based view of the universe of knowledge was common within both information science and the wider academy” (Szostak, Gnoli, and López-Huertas 2016, 1), interdisciplinary researchers must navigate an array of information resources that have tended to become fragmented and specialized, each often with its own distinct controlled vocabulary. In this environment, interdisciplinary searching becomes “an arduous undertaking for the end-user” (McCulloch 2004, 298).

Given this context, in this paper, consideration of interdisciplinary searching as a use case for vocabulary mapping includes: (1) discussion of the challenges posed by interdisciplinary searching and the need for continuing evolution of discovery tools; (2) analysis of the guidance provided by ISO 25964-2 in creating vocabulary mappings that could aid in interdisciplinary searching; (3) approaches to assessing the value added by vocabulary mapping from a user perspective.

2.0 Challenges of interdisciplinary searching

The most recent edition of the text Interdisciplinary Research: Process and Theory (Repko and Szostak 2020) includes a chapter on “Conducting the Literature Search” and observes that subject-oriented indexes and databases are of particular interest in this process. The reader is cautioned that each of these has its own thesaurus and is advised: “As researchers move from discipline to discipline in search of different insights on the same topic, they need to check each discipline’s thesaurus to find the term(s) to search for” (136).

The recognition of the need to support interdisciplinary searching is not new (see, for example, Smith, 1974; Weisgerber, 1993), but there have not been a large number of studies of interdisciplinary information seeking reported in the literature. Some key findings of completed studies include:

Bates (1996): “scholars in interdisciplinary fields may have to engage in both substantially more information seeking—and of a different kind—than scholars in a conventional discipline” (159). Information scatter contributes to the problems faced by interdisciplinary researchers.

Palmer (1996): “information probing” is an important type of information work for interdisciplinary researchers. “Probing is investigative in nature and takes place outside of the scientist’s core knowledge domain” (169). With each new domain, there are new terms and concepts to learn and analytical approaches to understand.
Spanner (2001): “One of the most frequently cited problems in crossing over into other disciplines is the problem of acculturation to non-affiliate disciplines, particularly in adjusting to conflicting vocabularies” (355).

In summary, the challenges confronting those engaged in interdisciplinary searching include the scatter of potentially relevant literature, some of which is outside the searcher’s core knowledge domain. More databases have to be searched, which involves having to navigate conflicting vocabularies.

3.0 Vocabulary mapping

Given the challenges posed by interdisciplinary searching, there is a strong interest in mechanisms to achieve subject or semantic interoperability (Zeng 2019). Dextre Clarke and Zeng (2012, 22) provide a timeline of landmark thesaurus standards in the English language, culminating in a new two-part international standard (ISO 2011; ISO 2013) developed by a working group with members from 15 countries, a chairman from the United Kingdom, and a Secretariat run by the National Information Standards Organization (NISO) in the US. As the title of this international standard, *Thesauri and Interoperability with Other Vocabularies*, makes clear, there is a strong focus on interoperability, “the goal of taking vocabularies which in most cases were intended to stand alone and relating them to each other in sufficient detail to permit searches drawn from one vocabulary to be effective in another” (NISO 2005, 82).

The Basel Register of Thesauri, Ontologies & Classifications (http://bartoc.org/) makes visible the wide range of extant thesauri, listing 766 ranging from *The Art and Architecture Thesaurus* to the *Zoological Record Thesaurus*. Where such thesauri are used as indexing vocabularies for specific discipline-oriented databases, a mapping can allow translation of a search statement from its initial formulation in one vocabulary to an equivalent statement in other vocabularies, where such correspondences exist. If vocabularies have only partial overlap in subject scope, mappings can be established only for those concepts covered in common, simplifying the task of mapping. Creating such mappings enhances the possibilities that formerly disconnected databases will be used in combination, especially to explore interdisciplinary topics. The goal is to enhance the ability of interdisciplinary researchers both to find what they want and to discover related information that they would not have known to look for.

As Zeng (2019) explains, the principles and practice of mapping are the prime focus of ISO part 2 (2013). The scope includes interoperability of thesauri with each other as well as with classification schemes, taxonomies, subject heading schemes, ontologies, terminologies, name authority lists, and synonym rings. Mapping establishes relationships between the concepts of one vocabulary and those of another, but challenges arise because vocabularies can differ with regard to structure, domain, language, or granularity.

Turning to the standard itself, Clause 5 *Objectives and identification* explains that the purpose of mapping is “to enable an expression formulated using one vocabulary to be converted to (or supplemented by) a corresponding expression in one or more other vocabularies” (16). When overlap between two vocabularies is small, selective mapping is carried out (Clause 6.5, 19). Three categories of mappings are distinguished: Equivalence mappings (Clause 8), Hierarchical mappings (Clause 9), and Associative mappings (Clause 10).
**Equivalence mappings** (EQ) encompass simple equivalence, 1 to 1 matching of concepts found in two or more different vocabularies (e.g., mobile phones EQ cell phones), and compound equivalence (1 to many). In compound equivalence a preferred term in one vocabulary may be represented in another vocabulary by a combination of two or more concepts/terms. The standard distinguishes between intersecting compound equivalence (e.g., genetically modified wheat EQ genetic modification + wheat) or cumulative compound equivalence (e.g., inland waterways EQ rivers | canals). In the case of exact equivalence, terms may be used interchangeably. In Clause 11 the standard also makes provision for identification of inexact equivalence (~EQ) when the most closely matching concepts in two or more vocabularies are not exactly the same. Such a mapping may lead to additional relevant items without too many irrelevant ones.

**Hierarchical mappings** (BM Broader Mapping/NM Narrower Mapping) between concepts are established when one is clearly broader than the other (e.g., rats BM rodents; rodents NM rats).

**Associative mappings** (RM Related Mapping) between concepts indicate situations where concepts do not qualify for equivalence or hierarchical mappings, but are semantically associated to such an extent that documents indexed with the one are likely to be relevant in a search for the other (e.g., e-learning RM distance education).

Clause 14 identifies how such mappings are accomplished: “Traditionally the identification of mappings is an intellectual process. It needs one or more experts familiar with the relevant subject field(s), fluent in the language(s) of the vocabularies to be mapped, and having a good understanding of the structure and conventions of the vocabularies” (38). Computer assistance may be used to automate the process in part by employing a matching algorithm and presenting candidate mappings for review by an expert. Because thesauri may continue to be revised and updated, Clause 15.3 highlights the maintenance that is needed if there are changes that affect the validity of the mapping.

Once created, Clause 12 discusses the uses of mappings in information retrieval as part of the indexing process or at the time of search. In interdisciplinary searching, mappings could be applied automatically in extending a search from one database to another, or human mediation could be used to select among options for search expansion. At a minimum equivalence relations enable translation of search terms from one thesaurus into those of another. Inexact equivalence, as well as hierarchical and associative mappings, offer possible avenues for search expansion.

Clause 16 discusses displays of mapped vocabularies, noting that the standard “does not seek to constrain the presentation of mappings data to end users” (45). The user need not be aware of the mapping process to make use of displays. For example, mapped terms “may be presented in the style of a tag cloud...without explicit designation of the type of mapping in each case.” (45).

**4.0 Evaluation of mapping**

Mappings connecting concepts in two or more thesauri will vary in the proportion of concepts involved, depending on the subject scope of each thesaurus. But even selective mappings require intellectual effort (and cost) to develop and maintain. As Kemp (2018, 82) notes, if metadata creators do not understand how metadata is used, it can be difficult
to make the case for enriching it. Research by Hider, Mitchell, and Parkes (2019) investigated the “retrieval power” added by subject indexing when compared to searching free text in a database and concluded that professional indexing using controlled vocabularies enhanced the yield of relevant resources. This approach can be extended to consider the retrieval power added by extending a search from the subject indexing employed by one database to the subject indexing via mapping in one or more additional databases when conducting an interdisciplinary search. The effectiveness of the mapping in retrieval can best be measured by submitting it to tests in the form of subject searches since its intent is to facilitate improved response to such queries. Such tests typically use measures of recall and precision based on relevance assessments of retrieved documents to gauge performance.

One exemplary study in this vein was carried out by Mayr and Petras (2009), who sought to assess the performance of a German Federal Ministry for Education and Research terminology mapping project, creating selective mappings (“cross-concordances” in their terminology) among multiple controlled vocabularies. Noting that while many mapping projects are undertaken, “the actual effectiveness and usefulness of the project outcomes is rarely evaluated stringently” (Mayr and Petras, 2009, 47), so they sought to determine how effective and helpful the mappings are in actual searches. Their experiments show the positive effects of mappings for search in heterogeneous databases, with interdisciplinary mappings having a higher positive impact on search results (Mayr and Petras 2009, 51). An early study by Smith (1974) demonstrated that using a mapping to extend searches to other databases beyond the major medical database for topics falling within nuclear medicine, bioengineering and computer applications, and physiology and biophysics yielded an increase in the variety of potential relevant resources by including technical reports, conference proceedings, and additional journals. In his expansive review of the role of thesauri in new information environments, Shiri (2012) found that study results demonstrate the usefulness of thesauri in both providing users with alternative search terms for query expansion and improving retrieval performance.

Sunny and Angadi (2018) sought to systematically review the literature evaluating the effectiveness of thesauri in digital information retrieval systems. In addition to studies that reported positive effects of an online thesaurus on retrieval performance, they found several studies focused on the usability of online thesauri. These studies reported positive reactions of participants in terms of identification and use of thesaurus terms; use in searching, browsing and navigation; and ease of use and helpfulness (65).

While evaluations of mappings between thesauri have been limited in number to date, the studies of single thesauri suggest needed directions for evaluation of mappings as they are developed and implemented, addressing both retrieval performance and usability. For interdisciplinary searching, semantic interoperability via mappings not only increases the success chances for distributed searches over databases with different thesauri, but it also can provide a view of a different disciplinary framework and domain-specific language, if the mapped vocabularies are made visible. Precision and recall of retrieval could be enhanced if mappings are differentiated using all the types of mapping described in Clauses 8 to 10 and if degrees of equivalence are marked as recommended in Clause 11. Especially where a given concept has no exact equivalent, provision of
inexact, broader, narrower, and associative mappings could be helpful in enabling selection of the best option for cross-browsing and cross-searching of various databases. In such query expansion, the user’s initial query statement is enhanced by additional search terms in order to improve retrieval performance.

In their discussion of new approaches to interdisciplinary knowledge organization, Szostak et al. (2016, 216) outline an ambitious research agenda to answer such questions as: What success do users have with the sorts of complex queries that interdisciplinary users often have? What about queries that members of one group might make about the practices or beliefs of other groups or disciplines? Evaluations of mappings can assess their efficacy in such situations.

5.0 Conclusion

The need for support for interdisciplinary searching is clear: “Search and communication absorb a great deal of the interdisciplinary researcher’s time. And failure to identify relevant information limits scholarly discovery” (Szostak et al. 2016, 220). Now that standards have formalized approaches to achieving semantic interoperability through mapping thesauri, next steps are implementation of more mappings and evaluation of their efficacy and usability. Echoing Szostak et al. (2016, 220), the goals of these efforts are to:

- Facilitate interdisciplinary searches by interdisciplinary scholars and students, both when they know what they are looking for and when they are seeking novel connections.
- Clarify terminology across disciplines.
- Facilitate the communication of research results to all relevant audiences.

As more thesaurus-enhanced user interfaces are implemented as search and browsing tools in a broad range of systems (Shiri 2012), from bibliographic and full-text databases to digital libraries, portals, open archives, subject gateways, and linked data repositories, the potential benefits of embedding mappings will increase. To take an example related to the conference theme of Knowledge Organization at the Interface, consider the researcher who is exploring approaches to enhance access to the Web for users with disabilities. Possible databases within scope include Inspec and Library and Information Science Source. In Inspec one finds the terms “handicapped aids” and “user interfaces”. In Library and Information Science Source one finds the terms “assistive computer technology” and “user interfaces (computer systems)”. A mapping of concepts common to Inspec and Library and Information Science Source would facilitate exploration of the computer science and engineering as well as library and information science perspectives on this topic. Extending the mapping to search PubMed via Medical Subject Headings could provide links to “communication aids for disabled” and “user-computer interface”. Given such a mapping, an evaluation study could involve the researcher in assessing items retrieved in response to a search request mapped across the three thesauri/databases for relevance. A mapping that included hierarchical and associative relationships that could be viewed by the researcher for possible search refinement or expansion (e.g., in Library and Information Science Source, “assistive computer technology” has “computers & people with disabilities” as a BT and “accessible websites for people with disabilities” as an RT) could be assessed for usability as well as retrieval performance.
In “As We May Think” Bush (1945) envisioned the memex, an information system with “a mesh of associative trails”. In the digital Web-based environment of 2020, vocabulary mapping offers one approach to creating a mesh of associative trails among a broad range of systems employing the thesauri included in the mapping. While this holds promise of facilitating interdisciplinary research, more evaluation is needed in order to assess its full potential.

References
Kemp, Jennifer. 2018. ”Metadata and Discoverability: A Use Case Overview.” Information Services & Use 38: 81-84.


Classifications as Linked Open Data
Challenges and Opportunities

Abstract:
Linked Data (LD) as a web-based technology enables in principle the seamless, machine-supported integration, interplay and augmentation of all kinds of knowledge, into what has been labeled a huge knowledge graph. Despite decades of web technology and, more recently, the Linked Data approach, the task to fully exploit these new technologies in the public domain is only commencing. One specific challenge is to transfer techniques developed pre-web to order our knowledge into the realm of Linked (Open) Data. This paper illustrates two different models in which a general analytico-synthetic classification can be published and made available as linked data. In both cases, a linked data solution deals with the intricacies of a pre-coordinated indexing language. The Universal Decimal Classification (UDC) approach illustrates a more complex solution driven by the practical requirements that the linked data model is expected to fulfill in the bibliographic domain, and within the constraints of copyright protection. The Basic Concepts Classification (BCC) is a new classification with a novel approach to classification structure and syntax for which linked data is an important vehicle for increasing the scheme’s visibility and usability. The report on these two cases illustrate some of the challenges of the representation of KOSs as LD and the possibilities that analytico-synthetic and interdisciplinary/phenomenon-based systems present for the representation of knowledge using LD.

1.0 Introduction
There is much excitement about the introduction of formal systems of knowledge organization into the infrastructure of the Linked Data (LD) and especially Linked Open Data (LOD) cloud. Expectations are grounded in the fact that linked data connect phenomena with shared (controlled) vocabularies. In theory, meaningful links from specific points in the cloud-based knowledge graph to normalized concepts in formal classifications can help to strengthen a shared conceptual infrastructure—not simply meaningful semantics but also effective syndetic routing among concepts. This objective was the core research question of the “Digging Into the Knowledge Graph” research project.1

Szostak et al. (2018, 527-28) pointed out three major challenges comprised sorting concepts, translating across domains and publishing KOSs as LOD. The Universal Decimal Classification (UDC) and the Basic Concepts Classification (BCC) —one disciplinary and the other phenomenon-based—were chosen as case studies to explore what problems along the journey of making KOSs available as LOD. As Siebes at al. (2019)

1 Digging Into the Knowledge Graph. https://diggingintodata.org/awards/2016/project/digging-knowledge-graph.
detailed, the process of moving into the realm of LD is composed of stages of conceptual and technological explorations and decisions. Under the former fall questions such as what information to make available in a machine readable form, to which extent existing vocabularies should be re-used, and whether or how to already enrich your LD prior to publication. Under the latter we find questions such as which web domain to use, how to design the URL’s, but also how to guarantee stability over time and how to document provenance during possible editions or versions of the LD publication.

Special challenges arise from the formal representation of KOSs as LD which are at once semantic and logistical. Semantic issues arise due to terminological diversity in the unorchestrated, self-organized nature of the LD cloud itself. The job of linkage from specifically well-defined points in a classification to a potential of semantic relations in the cloud is a non-trivial research task. Methodologically, and when dealing with Linked Open data (LOD), different routes for interlinking exist: point-to-point explorations in the process of publishing a resource as LOD (Siebes et al. 2019); inspection of LOD clusters as literary warrant (Martinez-Avila et al. 2018, 10); and translation between knowledge domains (Eito-Brun 2018; Marcondes 2018).

For KOSs which come with an extended legacy (a long history of well curated editions), such as the UDC, the choice of the appropriate namespace is non-trivial. We report approaches taken to publish the UDC and BCC as LOD enabling seamless integration into the cloud. Problems tackled in the process encompass data modelling, design of applied web technology (e.g., URI design), versioning/instantiating, licensing, extending KOSs published as LOD, and other possibilities to disseminate, exploit, and enhance KOSs. Publication of a KOS as LD is not trivial; rather, it requires a whole process of which many parts need to be accomplished first off-line.

The task of translating a KOS into LOD is challenging in many ways. In the aforementioned conceptual stage, selectivity is one aspect seldom discussed. The first task is not to transfer the whole of the KOS to the new (resource description framework or RDF) data model, but to chose those parts of the KOS which are most importantly available in a machine-readable Linked Data format. In this process, it is recommended to use already existing vocabularies to express selected features from the KOS in the new data model. This translation may allow, or at least facilitate, a translator to translate KOS terminology into items already mapped into LOD schemas. A second task is one of mapping connections from one RDF schema to another. Both tasks are far from being a mechanical mapping process, but require research as exemplified further below.

More particularly, this paper illustrates two different models in which a general analytico-synthetic classification can be published and made available as linked data. In both cases, a linked data solution deals with the intricacies of a pre-coordinated indexing language. The UDC approach illustrates a more complex solution driven by the practical requirements that the classification linked data is expected to fulfil in the bibliographic domain, and within the constraints of copyright protection. The BCC, interdisciplinary in nature, is a new classification with a novel approach to classification structure and syntax for which linked data is an important vehicle for increasing the scheme’s visibility and usability.
2.0 The BCC linked data publishing model

The Basic Concepts Classification (BCC)\(^2\) was created by Rick Szostak for the purpose of providing structured direct access by phenomenon to documents (and the ideas expressed in them). The BCC grew by the addition of schedules of (mostly verb-like) relators and adjectival/adverbial properties added to the original schedule of phenomena. Documents (objects, ideas, concepts) can be expressed with combinations of phenomena, relators and properties, either in symbolic notation (classified form) or in natural-language sentence style.

The primary difficulty in mapping a universal (i.e. general) KOS such as the BCC to LOD is that the BCC is intended to be able to classify almost anything (See Szostak 2019 for an overview of the BCC). The LOD cloud is also universal in extent, but achieves this universality with millions of distinct terms of varying degrees of specificity. A perfect mapping of the BCC to LOD would be able to encompass the entire cloud, but only by expanding BCC classes to such an extent that they would cease to be useful for classificatory purposes. The translator is left with the task of selecting points in the LOD cloud which hopefully encompass as much related information as possible.

The first task of the translator is to understand the relations, overlap and accepted usage between the current LOD cloud schemas. The initial impression on the translator is a bewildering array of options, some new, growing and maintained (Wikidata, DBPedia, OWL, SKOS, FOAF) and others suffering from disuse, abandonment or deprecation (Freebase). This array of options is a strength of LOD, for anyone can say anything about any topic (this is the so-called AAA rule that governs the Semantic Web), but for the translator it is very daunting to try to figure out if someone else is trying to say the same thing as your KOS.

Within the BCC there are essentially nouns (phenomena) and verbs (relators) [there are also adjective-like Properties that can be treated in much the same way as nouns]. Phenomena are significantly simpler to map as the translator needs to choose a sufficiently large schema and map terms in the BCC directly to those matching entities. As an example we have mapped the phenomenon of Art (http://purl.org/basic/a-art), using the relation of sameAs from OWL (http://www.w3.org/2002/07/owl/sameAs), to the DBPedia entry on Art (http://dbpedia.org/resource/Art). This is a reasonable mapping, and it implies that anything anyone has classified as Art using DBPedia is also classified as Art in the BCC. Note that this implies to a graph query engine that the terms of the BCC and DBPedia are identical and can be merged. This property may not be ideal. In our example, within DBPedia a movie is not art, but rather it is a subclass of work (also it is identified as equivalent to schema.org movie where it is a creative work). In the BCC, film is indeed a subclass of art (http://purl.org/basic/ar2-film) meaning that the classification is now disjoint and films classified as Art in the BCC will map to DBPedia incorrectly (at least according to DBPedia's definition). This is a general problem, for a controlled vocabulary such as the BCC is generally of greater breadth than an uncontrolled vocabulary (LOD).

One of the first points that the translator needs to comprehend is that it is possible to indicate the cardinality of relationships within LOD. For example, within SKOS there are classifications for broader and narrower, where the former implies that the object of

\(^2\) https://sites.google.com/a/ualberta.ca/rick-szostak/research/basic-concepts-classification-web-version-2013
the triple is broader than the subject, and narrower implies the inverse. For BCC Art one might be tempted to say that \'(DBPedia:art SKOS:broader BCC:art\', which indicates that everything DBPedia considers art is art in the BCC, but not everything the BCC considers art is in DBpedia art. The downside of using broader and narrower is that the mapping of the reciprocal is ambiguous (there is no way to know if a BCC-Art object should be DBPedia Art). Further, there are likely examples mapped to DBPedia-Art that are not in BCC-Art. The true cardinality of the relation is that there is a significant amount of overlap between DBPedia-Art and BCC-Art, and therefore we reasonably consider them the same, given our goals. That is, we allow some small degree of inaccuracy in translation in order to indicate a broad overlap in meaning.

For the translation of relators, the task is compounded as relators are used in the BCC to tie phenomena to one another, but in a more lexical way than LOD. In the BCC, relators can be used in conjunction with phenomena to add specificity to the classification. In terms of LOD, the word “visual” could be represented as “by pictures (/T7p),” where “by” is a relator and pictures is a phenomenon, but the idea represented by the two terms is smaller in scope than either term together. The translator may want to consider the effect of mapping the word “by” to any other definition, as while they may appear similar, what this implies is that for an object already mapped in LOD to be mapped to BCC it would have to link to both items in some way (which is unlikely). Again here we were faced with a decision of imprecision and decided to create an independent classification for the relator “by.”

The goal of translating a KOS to LOD is not to achieve perfection, but rather to cast a broad enough net so that the first iteration of the KOS can bring in terms that are related closely enough to its topics to test whether the KOS is capable of their classification. To this end, we begin with accepting the imperfect and hoping that it allows for iterative improvement.

3.0 The UDC linked data publishing model

The UDC has been one of the most widely used KOSs in the bibliographic domain for over a century. It is often used in conjunction with and complementary to thesauri, subject heading systems, and special classifications. During its lifetime, the classification has undergone many changes and has been made available in many languages and versions. The current UDC data standard, the UDC Master Reference File (UDC MRF) has had over twenty updates released since 1993, with 50% of the current 72,000 set of classes having been added or changed. The UDC data also include 12,000 cancelled (deprecated) classes that redirect to new classes. The scheme is currently owned, maintained, and developed by an international consortium of publishers, on a self-funding and non-profit basis.

The UDC scheme organizes concepts and subjects within traditional forms of knowledge (disciplines) allowing concepts and classes of concepts to be freely combined both within and between subject fields to express any level of complexity that information resources present. When both classification schemes and bibliographic metadata are published as linked data and are connected, they form a complex and dynamic knowledge space that shows the ways we create, interact with or utilize information. Classmarks stored in millions of bibliographic records hold valuable information about the contents of these collections. Once UDC classmarks are linked back
to the classification scheme from which they originated, it is possible to capture their meaning and establish further meaningful associations within and between collections (c.f. Slavic 2017). These connections made through linked data can help to:

- enrich bibliographic data to support information discovery by increasing subject access points using UDC terminology, by enabling semantic expansion (broadening); and by improving precision through contextualization;
- improve systematic presentation, grouping, and visualisation of resources and collections (linear or multidimensional) to facilitate browsing and serendipitous discovery of information;
- link the classification to other KOS to enable cross-collection information discovery;
- validate and update local classification data and local authority files or bypass local and obsolete classification data in information exchange.

Apart from many practical aspects of interest, UDC linked data development represents a good testbed for further research especially through its interaction with other KOSs. As an example of an analytico-synthetic and faceted scheme, it provides a case study for managing the alignment between the simple codes that appear in the scheme and the complex classmarks generated through document indexing (that contain unlimited numbers of combinations of UDC classmarks).

3.1 Challenges and solutions

While longevity and widespread use represent strong arguments for sharing the UDC as linked data, this also requires more responsibilities and presents further difficulties. In 2011, an extract from UDC of 2,600 classes was published as linked open data (LOD) in SKOS format. This experiment proved to be a valuable experience. As more and more library catalogues became available as linked data, we learned about the magnitude of the incompatibilities between UDC classmarks in bibliographic records and the UDC standard data.

Library linked data (LLD) clouds that were observed contained specific and complex UDC classmarks that can only be resolved through the access to the complete UDC content. However, the main cause of mismatch between UDC namespace and LLD is in the fact that libraries continue to use deprecated UDC codes. Thus, it became clear that a UDC namespace has to include not only a complete content of the UDC MRF, but also a significant collection of historical data and concordances between cancelled and new classes. Needless to say, the UDC LD use from 2011-2019 indicated that programs utilizing the UDC namespace (or those creating them) have little awareness of the UDC data structure, semantics, syntax, provenance, versioning, and changes and may not be able to process and select UDC data accurately or make good use of them.

In order to serve its purpose in a bibliographic domain, the UDC namespace has to provide a robust solution for the linking and semantic alignment between classmarks in bibliographic records and those in the UDC LD cloud. This has to be achieved irrespectively of the fact that the classmark strings in library data include combinations of simple UDC codes or that some may be deprecated or generated through wrong local practice. In order to achieve this, important changes had to be made to the ways and format in which UDC LD is published. This included the change of the URI format and the change
of the RDF schema, but most importantly, instead of a UDC linked data-dump we opted for a more complex UDC Look-up service.

The main premise of the UDC linked data service is that it ought to support practical use of the scheme as well as to protect the UDC publishing in a way that its future is safeguarded. This specifically means that only a small part of the UDC data shall be published as Linked Open Data (LOD) and most of the UDC LD content would be licence protected, i.e., linked data (LD) “behind the barrier.” The UDC LD-based terminological service must support the following features:

1. Programmatic access to:
   a) One LOD set: the UDC Summary containing 3,000 classes (under CC BY-NC-ND 2.0 licence)
   b) Two LD sets behind a UDC MRF licence barrier: a) Abridged edition (12,000 classes), and b) UDC MRF (72,000 classes), including all twenty versions of the UDC MRF and historical data comprising 13,000 cancelled (depreciated) classes and their redirections to new classes;
2. A UDC Look-up service that: a) parses and resolves (interprets) a classmark originated from bibliographic data and links its components to relevant records in the RDF data store; and b) upon request supply URI(s) for UDC classmarks or the full RDF records.

The architecture of the UDC Look-up service has the following components:
1. RDF stores (three Virtuoso databases: the UDC Summary, the Abridged edition, and the UDC MRF) with SPARQL endpoints accessible only via a restricted RESTful API layer which uses pre-designed SPARQL templates for query execution.
2. Apache web server and custom written UDC parser written in PHP and Java. The Authentication process is handled by standard shared and private authentication keys. The HTTP/Get parameters and the HTTP headers inform the server about the type of desired result (e.g., HTML, RDF-Turtle, JSON).

Although the UDC Look-up service is planned primarily as an API for programmatic interaction it will also have an html interface for human interaction with the service. It is assumed that the API would be queried by programs submitting simple or complex UDC classmarks either to get correct URIs for UDC codes or to retrieve complete RDF records. The HTML interface allows humans to verify and explore the provided classmarks in which the parsetree, versions, and RDF translations are expressed. The most important part of this service is the “UDC interpreter”, i.e., a program that parses complex UDC strings. This interpreter is based on a series of algorithms developed in an earlier project by Attila Piros (cf. Piros 2017). The UDC notation system allows for 100% accuracy in parsing of UDC strings using several groups of algorithms. Figure 1 shows an HTML interface in which a complex UDC number is split into components that, in this case, of all them valid UDC classes. The service executes queries against the UDC Summary, the UDC Abridged edition or the UDC MRF and in the second step it generates an RDF representation of the information selected by the user/machine from the previous step. For clarity, the terms shown in bold underline font in Figure 1 are resolvable primitive UDC terms.
3.2 Steps in publishing UDC linked data

This section outlines some of the key decisions and steps in the UDC LD service design. They broadly follow the ten step guidelines described by Siebes et al. (2019).

Selection of data. An important effort in this project was put in the strategic thinking and planning of UDC linked data and in particular to do with the selection of data to be published. The UDC Summary, the UDC Abridged Edition, and the UDC MRF are maintained in different MySQL databases and the same set-up is replicated for the RDF store (three Virtuoso databases). The selection of these three datasets is based on the well-established practice in UDC data use and publishing. They are representative of two kinds of access to UDC data: open access and access through a UDC MRF licence requiring an authentication process based on authentication tokens (managed outside the service itself). With respect to the supported languages, the UDC Summary contains language data in 57 languages. However, in this phase the Abridged and MRF datasets are available only in English. UDC data comprise many data elements that are required for data management and publishing, for the linked data, we selected only 14 data elements. In terms of sequence of data release, the UDC Summary (the LOD set) was given priority due to the large community of users.

URI name strategy. The UDC namespace was already established in 2011 and will remain as udcdata.info. The UDC experience shows that the decisions regarding the URI are far from being trivial. In the 2011 LD version, we opted for URIs that had the format of the following example: “udcdata.info/068288” in which the number “068288” represented a UDC record identifier for the notation =162.3 Czech language. An important reason for not including, at the time, UDC notation in the URI was the practice of the occasional re-use of deprecated notations (usually after 10 or more years). Thus, notation on its own was considered an unreliable identifier. Once historical versions of the MRF are included in linked data, a version code can be use to contextualise the notation, so we opted for a structured URI that includes UDC notation in the following format: “udcdata.info/MRF93/162.3.” In this example, the element “MRF93” represents the earliest MRF version in which this UDC classmark appeared, i.e., the version in which it was introduced for the first time. The advantage of this approach is that it makes easier for libraries to generate classmark queries to be launched against the UDC Look-up service and allows for human control of URIs (should this be required). An inconvenience with this approach is that UDC classmarks contain symbols and punctuations.
which are encoded automatically as they get processed, thus udcdata.info/MRF93/=162.3 becomes udcdata.info/MRF93/%3D162. This change of the URI format means that a new service must contain the mapping between the old 2011-2019 URIs and the new URI systems.

Use scenarios, serialization and resolution of UDC codes and URIs. When it comes to the Linked Data serialization of the UDC data source, we have to consider various scenarios in which UDC namespace will be accessed. Since the service is primarily aimed for machine access, we need to have disambiguation mechanisms combined with a clear guidance to make the programmers aware of the various choices that apply. For example, often the only information libraries have about the UDC is the classmarks and the location of the UDC Look-up service. They are not aware of the UDC MRF versions, including whether classmarks contain valid or deprecated numbers or whether they have licence, i.e., authentication token, to query full UDC data. Their queries may have the following format “udcdata.info/681.3(035)” - UDC Look-up service will parse and resolve the query indicating that notation 681.3 is deprecated and replaced by 004 and may return a RDF statement with sets of URIs expressing the relationship between these two numbers. If later at time an entity (machine or human) without access key for this dataset tried to query these URI’s at the UDC namespace, the authentication layer would prevent this request from being executed and return a meaningful error message, eventually combined with some sparse information about the result of the query (e.g., a superclass which the concept shares both from the MRF version and the UDC-summary version).

Selection of RDF schema. Following the parsing stage, URIs for individual classmark components and their grouping are generated using RDFs. For the full RDF records we use the SKOS format as it is widely used in the KOS publishing community. Equally, we wanted to maintain continuity with the 2011-2019 UDC linked data version. Below we can see the current mapping between UDC MRF data elements\(^3\) and the SKOS schema, which is extended by UDC sub-elements (in italics):

<table>
<thead>
<tr>
<th>UDC number (notation)</th>
<th>skos:notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>class identifier</td>
<td>skos:Concept</td>
</tr>
<tr>
<td>broader class</td>
<td>skos:broaderen</td>
</tr>
<tr>
<td>caption</td>
<td>skos:prefLabel</td>
</tr>
<tr>
<td>including note</td>
<td>skos:note</td>
</tr>
<tr>
<td>application note</td>
<td>skos:note</td>
</tr>
<tr>
<td>scope note</td>
<td>skos:scopeNote</td>
</tr>
<tr>
<td>examples</td>
<td>skos:example</td>
</tr>
<tr>
<td>see also reference</td>
<td>skos:related</td>
</tr>
<tr>
<td>revision history</td>
<td>skos:historyNote</td>
</tr>
<tr>
<td>introduction date</td>
<td>skos:historyNote</td>
</tr>
<tr>
<td>cancellation date</td>
<td>skos:historyNote</td>
</tr>
<tr>
<td>replaced by</td>
<td>skos:historyNote</td>
</tr>
<tr>
<td>last revision data</td>
<td>skos:historyNote</td>
</tr>
</tbody>
</table>

\(^3\) The UDC MRF data elements schema is available at:
In the future, we plan to move towards more formalized schemas from the OWL stack. This would enable a precise formalization which allows semantic verification of classmark strings, the vocabulary itself (e.g., when new concepts with their constraining properties are added or removed in future releases), and rich inference via transitivity, reflexivity, etc.

**4.0 Conclusion and future work**

As in many LD projects, the planning phase of both the UDC and BCC linked data projects took more time than originally anticipated. What is often underestimated is that the translation or transference of a resource to another medium or another technology is not merely a technological enterprise but is in essence coupled to a variety of research problems. The process can be compared to the mapping of vocabularies to each other, which is also not a mere mechanical process but entails all kind of research and editorial decisions, which in turn will influence how a KOS resource is further used. To operate on the scale of the web and with in principle unlimited outreach and spreading, the problem is only augmented. For both the UDC and BCC, key decisions had to be made through the combination of expertise in LD technologies and publishing models, on one hand, and expertise in the UDC or BCC schemes, datasets, and publishing models, on the other. More time for reflection, research, learning and discussion than envisioned was necessary in all key stages of the project. UDC and BCC are KOSs of a different type. The BCC is newer, experimental and still growing structurally. The UDC is one of the few authoritative KOSs for bibliographic databases, implemented widely, and based on a long history and fully developed KO principles of further development and implementation. Hence, the requirements for the LD publication are very different, and combining them was not part of the DiKG project. In this paper, we describe the different challenges those two KOS are exposed to during the LD publication.

Planning and developing of the UDC namespace in the form of a Look-up service presented challenges primarily because it is both a new and a complex approach to KOS publishing, also in the realm of established semantic web practices. Web-supported access to the UDC for humans based on a multi-tier license access which combines free access to part of the resource with licensed access for experts. This needs to be mimicked in the LOD transition. In our approach, LOD and “LD behind the licence barrier” models of publishing are combined and involve 3 different levels of classification data aimed at different audience and use scenarios. An important part of the UDC LD cloud is its historical data that will hopefully enhance the usability of UDC in the bibliographic domain where historical and obsolete classification data appears frequently. The most novel and key function to the Look-up service is the UDC interpreter. The UDC namespace is envisaged as a one-stop shop for querying and validating UDC data and it also illustrates a more complex, but hopefully more robust, model of KOS publishing as linked data. This UDC namespace offers a good environment for linked data and library linked data study and research on KOS alignments and integration.

**References**


The Subject Dimension of Authorship
A New Perspective of Provenance in KO

Abstract:
Authorship is a highly thematic and complex element since it brings aspects related to the specificity and context of works’ production. Considering that contexts and subjects are dynamic and vary over time and space, this paper aims to present the concept of subject provenance stressing that authorship can be considered as a product of a context – subject historicity, specific themes, and interpersonal networks. The approach taken to identifying essences in this paper is inductive. The discussion highlights that when considering an author as a subject herself/himself and as an access point – from the concept of subject provenance – it is delineated an organic relationship between the authors who participate in the same epistemic community, enriching the representation in the scientific literature. Thus an invisible but strong ribbon that leads to the complete subject context identification ties the author and the content of the document.

1.0 Introduction
Authorship has been traditionally linked to descriptive representation – also called cataloging – in information organization. Fields related to descriptive representation of documents used to be understood as neutral, explicit, and easier to identify, however the concept of authorship could be expanded to improve access to information going beyond the identification of “the name of the author”.

Authorship is a highly thematic and complex element since it brings aspects related to the specificity and context of production situated in a space-temporal framework (Guimarães 2017). In Literature, for instance, authors themselves can be the subject of literary studies, and in Philosophy subjects whose genesis can derive from their authors (e.g. Platonism, Socratic school). However, this approach is still incipient in the LIS scientific literature whose subject representation does not always consider the authors’ contexts (theoretical influences, schools of thought, points of view etc.).

For this reason, the concept of subject provenance is especially important once it contextualizes the authorship as a product of a context, such as subject historicity, specific themes, and interpersonal networks. In other words, an author can be also considered a representative of an epistemic community, as “a network of professionals with recognized expertise and competence in a particular domain” (Haas 1992, 3); of an invisible college, as “a set of interacting scholars or scientists who share similar research interests concerning a subject specialty, who often produce relevant publications to this subject, and who communicate both formally or informally with one another” (Zucalla 2006, 155), or even of a knowledge domain, as “a group with an antological base that reveals an underlying teleology, a set of common hypotheses, epistemological consensus on methodological approaches and social semantics” (Smiraglia 2012, 114).

Our hypothesis is that the more specialized is a domain the more an author is considered as a subject herself/himself – or, sometimes, an specific approach to a subject, and a powerful subject point of access with “power to influence” (Wolfram 2016).
1.1 The aim and scope of the study

The concept of subject is a complex one whose perception, for subject cataloging purposes, depends, from one side, on the knowledge of cataloger “about the structure of the subject and, on the other side, about the nature of the contribution that the document is making to the advancement of knowledge within a particular discipline” (Rowley and Hartley 2008, 109).

Hjørland (2017) points out that the subjects of documents are not a simple concept and comprises different views: from “content oriented” to “request oriented” ones. In this sense, the author highlights that “different definitions or implicit views of ‘subject’ is connected to different approaches and paradigms in information science.” As a consequence, “the subjects of a document are its informative or epistemological potentials, that is its potential of informing users and advance the development of knowledge” (Hjørland 2017, 62).

It is exactly on the determination of the epistemological potential of the subject that the author arises as an important contribution, once his/her approach reflects, to a certain extent, issues that bestow historical, political, economic, and technological contingencies on humanity.

In this context, this theoretical research starts from studies carried out by Moulaison, Dykas and Budd (2014) and Szostak (2015a; 2015b) with the aim to propose a new perspective of provenance in KO (Guimarães and Tognoli 2015; Tennis 2016) for the representation of scientific literature. This issue goes beyond the traditional descriptive representation with the purpose of reaching the essence of subject representation. The approach taken to identifying essences in this paper is inductive.

2.0 Authorship and Authority Records: Perspectives to Subject Representation

Authority data represent controlled access points and other information that institutions use to collocate works by a specific person, family, or corporate body, or the various editions of a title. For that, headings and cross-references for names and titles are constructed. In the bibliographic universe, guidelines and conceptual models bring elements to information organization in libraries based on relationships between attributes and entities.

The discussion promoted in this paper can be understood in the conceptual perspective, when it was tried to shed light on the extension of catalogers’ lens toward subjects based on “who is the author” as a contribution to a better specificity in subject representation.

The construction of authority records regarding person, corporate body and family is an intellectual work whose results promote consistency and, thereafter, credibility to the catalog and the information system as a whole. Consequently, it furnishes better elements to establishing further relationships with other records.

For instance, a pseudonym is an element of cataloging which is registered as an authority access point. Milani and Sousa (2018) states that understanding the presence of pseudonyms in catalogs can support subject representation once they can offer elements to the contextualization of the subjects assigned to a document. The mentioned authors present some perspectives regarding subject representation of pseudonyms of women authors who chose or were forced to sign their works with male or ambiguous names, especially between the early nineteenth and mid-twentieth centuries.
Knowing that some author used or uses a pseudonym to sign her/his works can complement two moments of subject analysis for cataloging purposes: 1) cognitive strategies: during the reading of the document, the cataloger would analyze through predictions if there are evidences that the fact of the author assuming authorship of the work through a pseudonym would somehow “shape de meaning” for the subjects approached, or would offer a different perspective of them; and 2) metacognitive strategies: in the step of identification and selection of terms, the cataloger can assign descriptors or create relationships stressing that a pseudonym was used as intellectual responsibility of that work (Milani and Sousa 2018).

Metacognitive strategies, which are conscious actions facing a problem and an objective can be proceed through some models or guidelines. Orientations like the set of dimensions to authorital perspectives for classification presented by Szostak (2015b) can offer feasible steps to subject analysis as presented above.

“The ‘who?’ dimension is addressed both when identifying disciplinary affiliation and in embracing personality dimensions and emotions under the intuitive approach to ethics. Disciplinary affiliation also indicates ‘when?’ and ‘where?’ an author is situated, as does the treatment of tradition within ethics. […] The ‘what?’ dimension is captured in several places, notably method, epistemology, and rhetoric. ‘Why?’ is even more extensively addressed in discipline, theory, method, ideology, ethics, and epistemology. And ‘how?’ is dealt within discipline, theory, method, epistemology, and rhetoric. Though the coverage of who, where, and when is limited, key elements of each is covered. What, why, and how are arguably far more important to the purposes of authorial perspective” (Szostak 2015b, 504).

These dimensions can also be used to analyze sources that may bring data to compound authors’ authority records.

While thinking about what have been viewed as sources of authority in determining information about persons in libraries, Dobreski and Kwaśnik (2017, 660) noticed that “persons are represented and distinguished from each other through common elements such as names, dates, and titles; however, closer inspection reveals the emergence of a number of critical differences concerning seemingly similar elements, as well as the scope and goals of the various standards and their respective definitions of a ‘person’”.

Much beyond a simple set of names, an author represents – and is deeply influenced by – her/his affiliations, the authors who have influenced her/him, authors she/he has influenced, and her/his predominant fields of study along the time (Moulaison, Dykas and Budd 2014). In this sense, access points made up of elements related to authorship would appear to be excellent bridges between the users’ information needs and the documents in a collection (e.g., research on works produced by authors born at a certain time, who attended a certain place, who fought in a specific war, who have a specific sexual orientation etc.).

Specifically in subject representation, “[w]hile the particular causal relationship(s) addressed in a work are the key aspect of what a work is about, other important aspects include any theory or method or data that was employed, as well as the perspective or worldview of the author” (Szostak 2015a, 595).

Aspects such as theory and method applied, philosophical and disciplinary perspectives of authors’ work, and other space-temporal aspects related to them would bring more elements to the subject cataloging process and to users’ analysis of the relevance of the documents retrieved in information systems. Therefore, authorship comprises a subject dimension once the author is a representative of a certain epistemic community, invisible
college and/or knowledge domain, and, as a consequence permeated by certain epistemological and methodological approaches, theoretical influences, and dialogical relations with other ones. In other words, authorship is important to furnish contextual elements to subjects in information systems.

3.0 A perspective of provenance in KO for the representation of scientific literature

To comprehend the authorship as a product of a context is essential to determine the epistemic communities that are involved in the production of knowledge. According to Meyer and Molyneux-Hodgson (2010), an epistemic community apart from producing knowledge also produces knowledge producers, shaping individual and collective trajectories. Accordingly, knowing that its genesis is imperative to know the actors involved in it, which are their trajectories and with whom they are related to in order to contextualize the authorship.

The context discussion is present in archival studies since always. The most important archival principles are related to preserving the documents in their original context. Context can be defined as

“[t]he organizational, functional, and operational circumstances surrounding materials’ creation, receipt, storage, or use, and its relationship to other materials [...] Along with content and structure, context is one of the three fundamental aspects of a record” (Pearce-Moses 2005, 90).

Another term used to describe the document context is archival bond, understood as being the relationship between records related to the same activity.

Considering the authorship as a product of contexts and relationships being part of an epistemic community we believe it could be thought from the principles and concepts already consolidated in archival science, such as the principle of provenance – thinking it as subject provenance – and the concept of organic collection – a set of records resulted of the routine activities of its creator (Pearce-Moses 2005).

The principle of provenance has been deeply addressed by archival studies since 1881, when Prussia State Archives issued a regulation establishing the Provenienzprinzip – based on the principle of respect des fonds – according to which public records should be grouped following the administrative units that created them (Schellenberg 2006, 175). The concept became a milestone in archival practice and today is considered the basis for what Tognoli, Guimarães and Tennis (2013) have called archival knowledge organization.

In archival studies, the concept is responsible to reveal the relationship between records and their context of creation, bonding the structures, functions, and activities of the producing entity. The International Council on Archives (2007, 11) defines provenance as

“the relationships between records and the organizations or individuals that created, accumulated and/or maintained and use them in the conduct of personal or corporate activity. Provenance is also the relationship between records and the functions which generated the need of the records.”

However, this contextual support offered by provenance is not limited to archival documents and has also offering elements to other disciplines such as LIS and Computer Science. The concept has been applied to different domains (e.g. preservation of digital records, digital evidence, digital humanities, linked data, information organization, and subject cataloging) with different points of view in order to determine records trust, especially due to the increasing use of information and communication technology.
In LIS, Tennis (2016) advocates the use of provenance to understand semantic changes in classification and order of things over time, besides being considered also a key access point in information retrieval. In KO, the same author relates the concepts of provenance and ontogeny:

“If provenance is defined as the chronology of custody and context (in the physical world often signaled by physical location) of some material, then we can see how revisions of indexing languages could change the context of a concept. With the change in context, the concept may change its meaning, and it is the meaning of the concept, in relation to other concepts and the documents they index that we care about in knowledge organization” (Tennis 2016, 94).

According to Tognoli and Guimarães (2019, 565), the concept of provenance can be enlarged to reflections on knowledge organization systems to every access point which presupposes – and depends on – historicity and context. Still, an author as an access point is not only an intellectual creator of the document but also a representative of an institutional context and of a subject domain in which the author has a certain position.

The idea of seeing the authorship as a provenance itself is enriched by the contextual approach that links the authors with an organic bond on a specific subject or interest area.

This contextual approach is based on relationship, which is “an association between two or more entities or between two or more classes of entities. To specify a relationship, we must be able, first, to designate all the parties bound by the relationship and, second, to specify the nature of the relationship” (Green 2001, 3).

As an example of the provenance applied to KO, Tognoli and Guimarães (2019, 566) analyses Birger Hjørland as an access point whose provenance embodies institution, academic background, theoretical views, research themes, and also and most important, the organic relationships – theoretical dialogues, convergences, divergences – that he establishes with other researchers of the same epistemic community.

Another example can be furnished by a comparative analysis between the production spheres. Considering the two papers “Knowledge organization: Its scope and possibilities” (Dahlberg 1993) and “What is Knowledge Organization” (Hjørland 2008) it is possible to observe that although both of them are concerned on discussing conceptual matters of KO, they furnish different perspectives of such conceptualization. Dahlberg, in a more ontological approach, and Hjørland in a more epistemological one, because of their academic background, their historicity and their institutional contexts. In this sense, epistemic communities can be identified in such a way that a record on conceptual issues of KO by Dahlberg is more closely related to the literature for Claudio Gnoli, for instance, while a paper from Hjørland has a closer dialogue with Jens-Erik Mai.

4.0 Discussion

We consider that knowledge about the profile of the intellectual background of an author is mandatory for a deeper subject cataloging of her/his works by providing more specificity. This knowledge would allow not only more specific subject headings (by incorporating, for instance, points of view or approaches for a certain subject), but also would furnish tools for the subject analysis and a wider range of subject relationships.

Access points in authority records, following the Statement of International Cataloging Principles (IFLA 2016), include entity’s authorized name or title, entity identifiers, various names and variant forms of the name or title of the entity, as it is possible to see in the example below.
Olson’s authorized heading in Library of Congress Authorities, which was updated in 2001.04.27, and is available in this permanent link lccn.loc.gov/n2001034168, offer the following description.

**HEADING:** Olson, Hope A.

Observing the field 670 (Source data found), the single source used to extract elements for this entry was a book published in 2001 by the author.

In Wikipedia, which is a source generally consulted for element’s extraction to construct authorized heading records when the cataloging standard allows it, the entry “Hope A. Olson” available at en.wikipedia.org/wiki/HopeA.Olson says:

Hope A. Olson is Associate Dean and Professor at the School of Information Studies at the University of Wisconsin-Milwaukee. She was previously a professor at the University of Alberta. From 2000 to 2004 Olson was the editor-in-chief for Knowledge Organization and she currently serves on its editorial board. She also serves on the editorial board of the Journal of Library Metadata. Olson has authored or co-authored over thirty peer-reviewed articles and book chapters, and has published three books: Subject Analysis in Online Catalogs, 2nd ed., co-authored by John J. Boll (Libraries Unlimited, 2001); Information Sources in Women’s Studies and Feminism, editor (KG Saur, 2002); and The Power to Name: Locating the Limits of Subject Representation in Libraries (Kluwer Academic, 2002). She received a B.A. from Gustavus Adolphus College, an M.L.S. from the University of Toronto, and a Ph.D. (1996) from the University of Wisconsin-Madison. Olson’s research focuses on critical analysis of subject representations and classification systems. Using feminist, poststructural, and postcolonial perspectives, she examines the biases inherent in hierarchical organizational structures.

In spite of being well constructed, this entry is not updated. The last line brings important elements, which characterize Olson as an author in a more accurate way.

The influence of the works of Hope A. Olson is assessed by Wolfram (2016, 336) through an egocentric informetric analysis of her published works, which demonstrate that:

Hope Olson has built an international reputation as a scholar and educator over a professional and academic career that has spanned several decades. Of particular note are her contributions to subject representation and classification, inter-indexer consistency, feminist perspectives and research methods. The co-citation analyses, both from the citation identity and citation image perspectives, reveal a scholar who not only is influenced by researchers in a range of areas of LIS and other fields but whose research is also cited in many areas of LIS and other disciplines. The textual analysis revealed equally broad influences based on the noun phrases present in Hope’s work and those used by authors who have cited her work. Dr. Olson’s contributions demonstrate that one individual can indeed influence (or inspire) researchers within and across disciplines (Wolfram 2016, 336).

Martínez-Ávila and Beak (2016, 365) analyze the epistemic stances and research methods and techniques of the thirty-three journal articles that Hope Olson published during the period 1991-2015, and identify that:

Hope Olson has used and introduced several poststructuralist methods and critical theoretical frameworks that are a fundamental part of her legacy in KO.

These studies bring important elements regarding the scholar Hope A. Olson, which can be part of her authority record in order to support cataloging and also the subject analysis.
process. The Wikipedia’s entry or Olson’s book from 2001 (used in her LC Authority record) do not convey important elements that compound her identity and power of influence.

It is important to sign out that the assumption of the name of an author as an element for a higher specificity in the subject representation of her/his works brings a special contribution to the building of a wider range of related terms. For this, two levels of thematic relationships can happen: subject to subject (in a more specific way) and author to author (both considered as a subject representatives).

For this, the book “The power to name: locating the limits of subject representation in libraries” published by Hope Olson in 2002 will be analysed. This book is considered one of the theoretical landmarks for a cultural approach in subject representation in KO.

The mentioned book is indexed under two subjects headings at the Library of Congress catalog: “subject headings” and “subject cataloguing.” Some other specific subjects, based on the author’s provenance and according to the subject specificity of the book would be: Deconstructive theory, Poststructuralism, Feminist theory, Gender and Information organization (or Knowledge organization), Race and Information organization, Ethnicity and Information organization, Culture and Information organization, Dewey Decimal Classification etc.

The assumption that an author has also subject nature can be shown by possible related terms that could be associated to “Olson, Hope A”. Once they share some common conceptions or approaches and whose additional search would be valuable when retrieving the book “The power to name”, such as: Beghtol, Clare; Berman, Sanford; Campbell, D. Grant; Frohmann, Bernd; García Gutiérrez, Antonio; Hudon, Michèle; Mai, Jens-Erik; Pinho, Fabio Assis; Tennis, Joseph T. etc.

Summarizing, the assumption that the author Hope A. Olson herself brings a thematic dimension for a deeper subject cataloging of her book “The power to name” allows us to identify a wider range of subjects to be assigned to the book: Subject Headings; Subject Cataloguing; Information organization – Social aspects; Deconstructive theory; Poststructuralism; Feminist theory; Gender and Information organization; Race and Information organization; Ethnicity and Information organization; Culture and Information organization; and Dewey Decimal Classification. Such approach would also allow us to relate her name to other authors that share common perspectives.

5.0 Conclusion

It is possible to conclude that authorship is part of subject cataloging once it brings elements to determine the subject context of production as well as the relationships that can be established with the epistemic communities which the author belongs to.

As Dobreski and Kwasnik (2017) say, “[l]ibraries are not cataloging people; they are cataloging identities” in such a way that the author and the subjects of the document are tied by an invisible but strong ribbon that leads to the complete subject context identification. In such context, the author, in that specific document, acts as a consequence of a time and Cartesian frame space (Guimarães 2017) determined by her/his ideas, point of views, and institutional context.

Studies that investigate some researcher’s attributes involving her/his epistemic stances, main theoretical frameworks, research methods and techniques, themes of research, citation identity (reflecting the influences of a researcher’s work, as explained
by Wolfram 2016), citation maps reflecting who cited her/him and who is cited by her/him etc. can be taken into account while thinking about an author as a subject.

Thus, when considering the author as a subject herself/himself and as an access point – from the concept of subject provenance – an organic relationship is established between the authors who participate in the same epistemic community, enriching the representation in the scientific literature.

References


© Ergon - ein Verlag in der Nomos Verlagsgesellschaft
Representing Concepts through Description Logic Expressions for Knowledge Organization System (KOS) Mapping

Abstract:
Mapping among KOS can be achieved by representing the concepts in each KOS through a canonical expression:

<table>
<thead>
<tr>
<th>KOS</th>
<th>Canonical expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDC</td>
<td></td>
</tr>
<tr>
<td>GND</td>
<td></td>
</tr>
</tbody>
</table>

We explored representing concepts through description logic expressions using a sample of 150 library KOS classes and subject headings, 50 each from the Dewey Decimal Classification (DDC), the Regensburg Verbundklassifikation (RVK), and the German Integrated Authority File (GND). Working from the ground up, we compiled a small vocabulary of relationships/roles and the beginnings of a faceted classification of elemental concepts. Large-scale application of this approach requires a large hierarchically structured vocabulary of relationships/roles and a universal faceted classification. We discuss the task of developing these tools drawing on many sources.

1.0 Introduction
A description logic expression is a combination of elemental concepts giving for each elemental concept the role it plays in the context. Adding the roles is a refinement of simple semantic factoring. A DL expression is a formal definition of a concept (Baader et al. 2017). We use a simplified version of DL expressions, similar to the Semantic Code (Perry and Kent 1958). On the other hand, we found it necessary to allow for nesting in DL expressions, indicated by [ ]. Some examples:

<table>
<thead>
<tr>
<th>KOS</th>
<th>Canonical expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>RVK:</td>
<td>Technology &gt; Traffic &gt; Traffic safety = &lt;isa&gt; safety ⊓ &lt;inEnvironment&gt; Transportation</td>
</tr>
<tr>
<td>GND:</td>
<td>Verkehrssicherheit (Traffic safety) = &lt;isa&gt; safety ⊓ &lt;inEnvironment&gt; Transportation</td>
</tr>
<tr>
<td>RVK:</td>
<td>PN808 Law &gt; … &gt; Hazardous material = &lt;isa&gt; law (topical area) ⊓ &lt;appliesTo&gt; [ &lt;isa&gt; goods ⊓ &lt;hasCharacteristic&gt; hazardous ]</td>
</tr>
<tr>
<td>DDC:</td>
<td>343.0938 Law &gt; … &gt; traffic safety = &lt;isa&gt; law (topical area) ⊓ &lt;appliesTo&gt; [ &lt;isa&gt; safety ⊓ &lt;inEnvironment&gt; Transportation ]</td>
</tr>
<tr>
<td>DDC:</td>
<td>343.09322 Law &gt; Transportation &gt; Transportation of goods &gt; Hazardous material = &lt;isa&gt; law (topical area) ⊓ &lt;appliesTo&gt; [ &lt;isa&gt; transportation system ⊓ &lt;objectTransported&gt; [ &lt;isa&gt; goods ⊓ &lt;hasCharacteristic&gt; hazardous ] ]</td>
</tr>
<tr>
<td>GND:</td>
<td>Gefahrgutbeforderungsrecht (Law on transportation of hazardous goods) = &lt;isa&gt; law (topical area) ⊓ &lt;appliesTo&gt; [ &lt;isa&gt; transportation system ⊓ &lt;objectTransported&gt; [ &lt;isa&gt; goods ⊓ &lt;hasCharacteristic&gt; hazardous ] ]</td>
</tr>
</tbody>
</table>

Classes or subject headings from two KOS that have the same DL expression refer to the same concept and can be mapped with skos:directMatch. A reasoner working on
a database of DL expressions can infer other relationships between concepts. For example, there is an associative relationship between RVK: PN808 and DDC: 343.0938, since Safety and hazardous are related; see Section 2.2.

For the context of this work see Balakrishnan et al. 2018, 2019. For the theoretical basis see Soergel 1972, 2011, 2017.

2.0 Creating DL expressions and analyzing pair-wise mappings

2.1 Creating DL expressions for classes / subject headings - The challenges

Requires a good understanding of KOS structure and considerable domain knowledge.

<table>
<thead>
<tr>
<th>Understanding KOS structure. Consider a class in its hierarchical context</th>
</tr>
</thead>
<tbody>
<tr>
<td>RVK: CM 5000 Information theory, cybernetics is actually</td>
</tr>
<tr>
<td>RVK: CM 5000 Psychology &gt; General, history and methods &gt; Information theory, cybernetics</td>
</tr>
<tr>
<td>= &lt;isa&gt;Information theory, cybernetics ⊓ &lt;usedAsMethodIn&gt; psychology</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Domain Knowledge is essential; often need to look up a definition.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A DL expression is crystallized domain knowledge</td>
</tr>
<tr>
<td>DDC: 150.1 Psychology &gt; philosophy, theory, systems, schools = &lt;isa&gt; philosophy, theory, systems, schools ⊓ &lt;studies&gt; psychology</td>
</tr>
<tr>
<td>GND: Philosophical Psychology = &lt;isa&gt; psychology ⊓ &lt;studies&gt; topics in philosophical psychology ⊓ &lt;usesMethod&gt; philosophical method</td>
</tr>
<tr>
<td>GND: Historical psychology = &lt;isa&gt;psychology ⊓ &lt;studies&gt; [&lt;isa&gt;topics in psych. ⊓ &lt;inTimePeriod&gt; past]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ambiguity: Caption is ambiguous, perhaps used both ways. Two DL expressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDC: 610. 82 Women in Medicine#1 = &lt;isa&gt; Person ⊓ &lt;sex&gt; female ⊓ &lt;agentIn&gt; Medicine</td>
</tr>
<tr>
<td>DDC: 610. 82 Women in Medicine#2 = &lt;isa&gt; Person ⊓ &lt;sex&gt; female ⊓ &lt;patientIn&gt; Medicine</td>
</tr>
<tr>
<td>GND: Ethnomathematics#1 = &lt;isa&gt; math. ⊓ &lt;practicedAmong&gt; identifiable cultural group</td>
</tr>
<tr>
<td>GND: Ethnomathematics#2 = &lt;isa&gt; curriculum subject ⊓ &lt;studies&gt; [ &lt;rel&gt; culture, &lt;rel&gt; math.]</td>
</tr>
</tbody>
</table>

2.2 Analysis of mapping pairs based on DL Expressions

Given the DL-Expressions for each class or subject heading in a pair taken from KOS A and KOS B, a system can infer the type of mapping, as shown in the examples.

<table>
<thead>
<tr>
<th>skos:exactMatch</th>
</tr>
</thead>
<tbody>
<tr>
<td>RVK: CP 3200 General psychology &gt; Feelings, emotion = &lt;isa&gt; emotion</td>
</tr>
<tr>
<td>GND: Feeling, emotion = &lt;isa&gt; emotion</td>
</tr>
<tr>
<td>DDC: 150.9#1 Psychology &gt; History, [biographic treatment, biography] &gt;</td>
</tr>
<tr>
<td>= &lt;isa&gt; history ⊓ &lt;studies&gt; psychology</td>
</tr>
<tr>
<td>RVK: CM 2000 Psychology &gt; History of psychology = &lt;isa&gt; history ⊓ &lt;studies&gt; psychology</td>
</tr>
</tbody>
</table>

The history part of the DDC class (without biography) is an exact match for the RVK class.
3.0 Toward a system relationships/roles and a universal faceted classification

To apply the approach illustrated at large scale requires a large list of relationships/roles and a large universal faceted classification. The following sections give a flavor of what needs to be done, but the task is monumental.

3.1 Relationship types / roles

For this exploration we introduced relationship types/roles as needed. Some are quite obvious, such as \(<\textit{isPartOf}>, \langle\textit{produces}\rangle, \langle\textit{hasCharacteristic}\rangle\), but others are not so common or quite specialized, such as the following: \langle\textit{usesMethod}\rangle, \langle\textit{usedAsMethodIn}\rangle, \langle\textit{studies}\rangle, \langle\textit{appliesTo}\rangle, \langle\textit{inEnvironment}\rangle, \langle\textit{agentIn}\rangle, \langle\textit{patientIn}\rangle, \langle\textit{objectTransported}\rangle, \langle\textit{atEducationLevel}\rangle.

Our ultimate goal is either to locate each relationship/role in a standard or widely used ontology or to contribute to some inventory of relationship types / roles.
3.2 A universal faceted classification of elemental concept

At this stage of the project, we introduced elemental concepts as needed for the DL expressions, standardizing terminology in this small set of 150 classes and subject headings. The next step would be to develop a universal faceted classification, a monumental task using many sources, including standard classifications, see Figure 1.

The nature of the classes in DDC and RVK often requires elemental concepts that are defined using or such as

- Comparison or harmonization (with NT Comparison and NT Harmonization)
- Cognition or intelligence
- Philosophy or theory or discussion of viewpoints or schools (frequent subdivision in DDC)

| We need Level of education, as seen from some concepts we used as components. The International Standard Classification of Education (ISCED) maintained by UNESCO includes a classification for Level of education. For purpose of illustration we built a hierarchy consisting mainly of ISCED concepts (ISCED does not present these as a hierarchy). We added a number of concepts, some to group ISCED concepts, some extensions further down (Kindergarten), because we know or expect that these are needed in building DL expressions for DDC and RVK. | ISCED level 0 or 1 Early childhood education to Primary ed.
  - ISCED: ISCED level 0 – Early childhood education
  - ISCED: ISCED level 1 – Primary education
- Secondary education
  - ISCED: ISCED level 2 – Lower secondary education
  - ISCED: ISCED level 3 – Upper secondary education
- ISCED: ISCED level 4 – Post-secondary non-tertiary ed.
- ISCED: Tertiary education
  - ISCED: ISCED level 5 – Short-cycle tertiary education
  - ISCED: ISCED level 6 – Bachelor’s or equivalent level
  - ISCED: ISCED level 7 – Master’s or equivalent level
  - ISCED: ISCED level 8 – Doctoral or equivalent level
- K-12 education
  - NT Kindergarten
  - NT ISCED level 1 – Primary education
  - NT Secondary education
- Youth and adult education level
  - Youth education level
  - NT Secondary education
  - Adult education level
    - NT ISCED: ISCED level 4 – Post-sec. non-tertiary ed.
    - NT ISCED: Tertiary education |

Figure 1. Hierarchy of Level of education

3.3 Concept definitions

Elemental concepts must be defined, particularly concepts that have been introduced specifically for DL expressions. Consider our pattern for DL expressions for subject disciplines:

- environmental science = <isa> subject discipline ⊓ <studies> environment
  - But what to do with psychology? We could come up with psychology = <isa> subject discipline ⊓ <studies> mental states and processes and behaviors
  - But that is a cryptic characterization of what psychologists study. Better:
    - psychology = <isa> subject discipline ⊓ <studies> topics in psychology

Now, we can create a full definition of topics in psychology and put many narrower terms under it, for example from (www.apa.org).

For another example, consider several related meanings of law:
law (subject discipline), law (topic area), and Body of law, statutes.

4.0 Conclusions and future work

In our exploration representing KOS concepts through simple DL expressions worked generally well. Comparing KOS mapping using DL expressions with mappings found in our database convinced us that it is worthwhile to test the idea further in a larger pilot that would use proper software to support developing and maintaining the systems discussed in Section 3 and to partially automate the creation of DL expression using linguistic analysis of captions.¹

References


¹ The data sets created in this exploration (list of relationships/roles, partially structured list of elemental concepts, definitions, and DL expressions) are available from the authors.
Classification System for Knowledge Organization Literature (CSKOL) 
Its Update, a Pending Task?

Abstract: Dahlberg created the Classification System for Knowledge Organization Literature (CSKOL) to classify the Section Knowledge Organization Literature of the International Classification journal, later known as Knowledge Organization Journal (KOJ). Three editions of the CSKOL have been published in the KOJ (1985, 1993, 1999). Formally, the CSKOL is a classification system specialized in KO. It is also a decimal faceted classification scheme. The CSKOL has not experienced significant changes. Nevertheless, KO literature has been classified with this system for 45 years. Authors perceive KO as a dynamic and interdisciplinary field, with various research lines and trends, continuously incorporating new concepts. There is some tension between the evolution of KO and the relative immobility of the CSKOL schemes. This paper suggests a revision of the CSKOL to update the terminology, prepare a classification tutorial and publish a complete and up-to-date official version.

1.0 History and editions of the system


Three editions of the CSKOL have been published so far:

i) First edition: published in 1985 (ICJ volume 12, issue 3), under the title of Classification Literature Classification (CLC). The outline of the scheme is simple: it shows the development of nine classes (numbered from 1 to 9). In most classes, there are no subdivisions beyond the third digit.

ii) Second edition: published in 1993 (KOJ volume 20, issue 4). In this edition, – under its current title-, the system is an appendix to an article where Dahlberg laid the foundations for Knowledge Organization (KO) (Dahlberg 1993).

In this instance, the author presented three tables: a first summary with ten main classes, a second summary with one hundred divisions, and finally, the extended scheme, with the complete system. In that same article, Dahlberg published a summary introducing the one hundred major divisions of the Information Coding Classification (ICC), an autonomous, universal classification system she created (Dahlberg 2008), and conceived as a complement to the CSKOL.

The ICC allows defining many subjects specific to KO by subdividing those subjects by discipline. For instance, it is possible to add to the class number 82 of the CSKOL (“Data Classing and Indexing”) the class number 24 of the ICC (“Electronics”), thus achieving the subject representation of information resources on indexing of documents on electronics (82-24).
When we compare the second edition with the first one, we find at least three differences: a) The addition of Class 0 (Form Divisions) to the schemes; b) The disappearance of some levels of specificity, due to the sole inclusion of topics up to the third digit; c) The addition of new topics under notations 158, 169, 199, 349, and 399, among others.

iii) Third edition: published in 1999 (KOJ volume 26 issue 4). This issue was entirely devoted to compiling the bibliography of the KOJ and the proceedings of international conferences, national conferences, and other ISKO publications. The entire bibliography was classified using the CSKOL schemes.

In a brief introduction, it is mentioned that the CSKOL is presented “as it is used by the present Literature editor of the KOJ [at that time, the specialist Gerhard Riesthuis], including some small changes made with a view to changes in the literature in the field of KO” (Riesthuis 1999, 192). No mention is made to the changes or the methodology used. One of the changes identified is the use of the asterisk (*) in those CSKOL class numbers that can be combined with ICC numbers.

2.0 Structure and principles of the system

Formally, the CSKOL is a classification system specialized in KO. It is also a facet-ed classification scheme, built from a top-down perspective. The consistent and exhaustive application of the facet analysis advocated by Ranganathan (1937) is noticeable throughout the CSKOL.

The CSKOL is also a decimal system, and, therefore, it has a hierarchical structure, similar to the DDC or the UDC classification systems: Class 0 corresponds to Form Divisions. Classes 1-9 are organized into three thematic groups that follow the sequence of the 'systematifier' (Dahlberg 1977).

The classification numbers have up to three digits, with a few exceptions where a fourth digit is added to some numbers. This allows organizing specialized collections with a specificity ranging from low to medium.

The CSKOL is currently divided into the following ten main classes:

0 – Form Division (form bibliographies, reviews, glossaries, textbooks, among others).
1 – Theoretical Foundations and General Problems.
2 – Classification Systems and Thesauri, Structure and Construction.
3 – Methodology of Classing and Indexing.
4 – On Universal Classification Systems and Thesauri.
5 – On Special Objects Classifications (Taxonomies).
6 – On Special Subjects Classifications and Thesauri.
7 – Knowledge Representation by Language and Terminology.
8 – Applied Classing and Indexing.
9 – KO Environment (professional and organizational problems, persons, and institutions in KO, policy, and legal questions, among others).

Dahlberg also gathered the nine thematic classes (1 to 9) in three groups:

Group 1-3: representing the constituent divisions of the field.
Group 4-6: representing the application of the constituent divisions of the field.
Group 7-9: representing the influence, application, and environment fields.
As for the terminology, expressions or phrases (generally pre-coordinated) prevail over terms or descriptors. For example: under 215- Characteristics and Kinds of Thesauri. No synonym relationships are established in the schemes; thus, the only terms included are authorized terms.

The system does not have auxiliary tables, though the decimal numbers can be complemented with the auxiliary tables of the Dewey Decimal Classification (DDC), the Universal Decimal Classification (UDC) or the Information Coding Classification (ICC). The CSKOL has no index: thus, it is necessary to look for the topics throughout the scheme. Besides, there is no classification tutorial we are aware of.

Although the CSKOL could be considered an ad hoc classification, created to organize the Literature Section of the KOJ, its use can be extended to classify collections specialized in KO in universities, research centres or private libraries belonging to researchers.

3.0 Primary approach to the system

KO literature has been systematically classified with the CSKOL for 45 years, without significant modifications to its schemes. However, authors perceive that KO is a dynamic and interdisciplinary field, with a great diversity of research lines and trends (McIlwaine 2003; Hjørland 2008; Ridenour and Smiraglia 2016) continuously incorporating new concepts. Therefore, there is some tension between the evolution of KO and the relative immobility of the CSKOL schemes.

However, the applicability of the system does not seem to have been in doubt so far, since the successive editors of the Literature Section (all of them well-known researchers) have continued using it. The amendments and modifications made are minor, as stated in various summaries of the Literature Section. These modifications do not appear in any known text, but they can be identified through the analysis of free numbers assigned to specific documents (such as the notation 949 vacant until 1999 and representing the term 'Authority control' since 2000).

The structure of the UDC was the model for the general analysis. The warrant was identified based on the list established by Barité (2019). The formal review of the CSKOL terminology was performed taking into account the principles of the British Standard BS 8723 (BSI 2005-2008).

In the first exploration, the analysis identified the following problematic issues:

a) The classification criteria used by the different editors have not been adequately spread. The only access way is through an inductive analysis of the system use over time.

b) There is no known protocol available establishing the guidelines for the revision of the schemes, and their periodic update.

c) It is necessary to add auxiliary tables (place, time, form, and others), or to expressly enable the combination with the corresponding auxiliary tables of other systems (DDC, UDC, ICC).

d) The lack of a general index of the system makes specific searches by subject difficult and is an important limitation for the analysis of the terminological coverage of the area, and the detection of gaps. It is also an obstacle to developing different research activities that need the terminology of the area as a basis.
e) The warrant used to justify the selection of the terminology is unknown; nevertheless, it is assumed that Dahlberg used her own expert opinion, a form of scholarly (or academic) warrant.

f) A complementary alternative would be to review the whole terminology of the CSKOL, in order to use subjects as descriptors. The general index could also be structured, thus becoming a list of KO descriptors.

There are two previous studies using the CSKOL: an analysis of the scientific production recorded in the Section KO Literature from 1991 to 1993 (Dahlberg 1995) and a doctoral thesis that evaluates the literary warrant resulting from this production between 1994 and 2009 (Barité 2011). In both cases, the emphasis was placed, on the production of quantitative and statistical data of scientific production in KO. Due to the methods used and the results obtained, they could be considered as two indirect ways of assessing the applicability of the system.

For example, Barité distributed the percentages of the 12,833 classifications made with the CSKOL between 1994 and 2009 in the ten classes of the system.

The classifications were distributed in four-year periods; to identify increasing or decreasing trends in academic production. Only 4 out of the 10 classes exceed the statistical 10% of classifications, three are around 10% (+/-2), and the remaining five classes are below these figures, with an unexpected value of zero literary warrant for class 5- Taxonomies (Barité 2011, 292).

This table shows the inadequacy of class 5, and the saturation of class 7 (23,3%). This may require new subdivisions or the transfer of some topics to other classes.

The use of the abovementioned sources (Dahlberg 1995; Barité 2011) and a recent one (Roszkowski 2020) may help to establish a methodology to examine the evolution of KO literature and to identify the areas of modulation and degrees of specialization of the domain (Tennis 2003), to make a new design of the CSKOL.

4.0 Conclusions

The CSKOL is a specialized classification system, created to organize by subject the documentation on KO. Therefore, it should be seen as a model system as far as its theoretical and methodological principles are concerned.

The preservation of a concept structure with minimal modifications over 45 years, while the subject field has experienced a strong change dynamic in the same period (new types of systems, new ideas, new methods, new products), compels to consider procedures of revision and update.

The two existing studies have been based on the literary warrant, as a methodology to identify - through scientific production over an extended period – the almost unused segments of the system, as opposed to others that are overused.

All these remarks might lead to justify the need for a partial revision and even a more general reformulation of the CSKOL system; this will surely imply a long-term challenge for the editorial team. Therefore, we suggest a revision of the CSKOL including the analysis of its terminology based on the literary warrant provided by the KO Literature Section, even through methodologies such as metric analysis, to eliminate obsolete terms, incorporate new ones, specify notations and turn the terms into descriptors; a tutorial on classification will also be required.
After completing the revision, the publishers should publish a complete and up-to-date official version of the CSKOL explaining the design and the classification criteria as guidance and reference for users.

References
Thiago Henrique Bragato Barros – Federal University of Rio Grande do Sul, Brazil

Touching from a Distance
Concept Theory and Archival Hierarchical Classification

Abstract:
This paper discusses the aspects and points of contact between the principle of archival provenance and concept theory to help to build better hierarchical classification schemes in archival theory and practice. Knowledge Organization, especially concerning classification, can help to construct better ways to represent and access archival materials. The aim is to systematize the relations between archival provenance and concept theory-based within KO and archival science. With these two fields, we can construct a methodological parallel and argued where and how concept theory KOS could help in the construction of better concepts to represent a given archival domain. We learned that yes, thinking about the terms - concepts - relating two approaches bring a more scientific background to the construction of classification, reducing in some sense, subjectivity, and bias in its development.

1.0 Introduction
This study aimed to discuss the possible relationships between archival theory and Knowledge organization in the spectrum of concept theory, especially concerning archival classification, which are constructed based in a functional-hierarchal classification.

We also seek, from the parallel, how one can help the other because we have a lot of complicated issues when we design archival classification schemes; we have concepts, principals, and approaches. However, we have difficulties when it comes to build, use, and reuse this scheme, so there is a need to develop better systems. More critical in light of newer digital records production systems.

2.0 The method
We aimed to construct a parallel and analyses between archival classification and concept theory. Considering that this is a short paper essay, we compared in a summarized manner how archival classification works in Hurley (1993), Sabourin (2001), ISO15489-1 (2016), Orr (2005), Foscarini (2006). Also, how archival classification theory relate to concepts and terms based on Dahlberg (1978; 1993) Hjørland (2008), among others, bringing concept theory to the archival classification discussion, we see significant methodological contributions for both KO and Archival Science.

3.0 Knowledge Organization and Concept theory
The field of Knowledge Organization can be understood as an area of studies, which encompasses the research, practice, and theory of a subdomain of Information Science historically focusing on bibliographic representation, constructing and evaluating semantic tools for information retrieval (IR). More specifically, KO is concerned with such activities as indexing, abstracting, and classifying bibliographical items within. According to Hjørland (2016), this field relates to “description, representation, registration, organization of documents, subjects, and concepts.” For that, some tools such as classification systems, lists of subject headings, and thesaurus are necessary.
We also can point out that KO investigates the nature and quality of knowledge organization processes (KOP) and the structure and function of knowledge organization systems (KOS). The role of (KOS) is to organize documents, representations of documents, and concepts (Hjørland 2008).

KO thus includes a wide array of research interests, including the theoretical basis of KOS, the history of KOS, terminological issues, domain studies, genre studies, and the social organization of the sciences, KOS are a sound basis for the development of information retrieval. We think there is a role that these systems can play concerning archival science, especially in archival classification were the Archival Knowledge Organization System can be built (Barros and Sousa 2019) but for most of the Archival Science history was not viewed as such (or even presented).

KOS can be universal as Dewey Decimal Classification (DDC) and Universal Decimal Classification (UDC) or as specialized as a taxonomy for a specific purpose, these specialized systems, and we can understand archival organization as a specialized system and provenance as a basis for the construction for classification schemes that intellectually reflect institutional domains.

For the development of KOS, concepts and terms are a basis for the construction of taxonomies, ontologies, or other systems. Concepts and terms are also the basis for archival classification; however, the final objective is different, one is to represent a scientific knowledge domain, the other is to represent an institution, business, and State records, all specifics knowledge domains. With that in mind, we think, Dahlberg offered the foundation of the term “concept.” Dahlberg proposes a concept system that consists of a triangular representation made up of reference, characteristic, and verbal form is a functional one. (Dahlberg 1978) The creation of reference occurs through three activities: prediction, denotation, and designation. The term “denotation” is synonymous with the reference step, while the term “predication,” represents the postulation of the reference. The term “designation” is the actual transformation of the reference and the characteristics of the concept to a verbal form. (Dahlberg 1978). We think this triangular relationship of terms and concepts building KOS, can help as secondary tools for archival classification.

4.0 Touching from a distance: Archival Classification and Concept theory

Classification is traditionally defended as a fundamental for the organization of archival materials. When we talk about archival classification, we are relating provenance or the origin of the record in linking with its contexts. When we compare these things, we are talking about function as a form to look for records, in a juridic administrative body. This classification delineates management processes, organization, and representation, and in the context of the archives, that is, the classification must have extreme representative power. (Barros and Sousa 2019). Besides the fact that it is a system of the organization deeply formalized and hierarchical by nature, it also has semantic aspects in its structuring and standardization possibilities. It is part of an interchanged process, but archival science hardly acknowledged that in its history or practice, and here is where concept theory and KOS enters, they are semantic tools as their bibliographical counterparts.
Eastwood (1994) and Duranti (1997) argue that only records together are archival records and evidence of the activities carried out by an institution, i.e., any document that is not organized by its function/activities, establishing a relationship with its origin (provenance) and its original order cannot be understood as an archival record. As RAD (Rules for Archival Description 2008, xxiii) dilates, “the principle of provenance means that the records created, accumulated or maintained by an individual or organization must be represented together, distinguishable from the records of any other creator.”

When we think about this, we can say yes, we understand, based on recordkeeping tradition, that provenance and contexts are the fundaments for organizing organic records. However, the way we do it is conceptual and semantic, but we lack tools and studies for acknowledging that. So, the system concerning its design is incomplete; there is room to work with ontologies or taxonomies, thinking of how we name things, not only why we call based on provenance and relate these instruments to who might use the record in any contexts. Basic methodology for archival management and classification—the functional analysis—began in the 1940s with Brooks (1940), Posner (1964), and it was systematized in a more “complete” way in Schellenberg (2003).

Concerning the process, Foscarini (2006, 41) established that we could define it as a preliminary investigation, followed by top-down functional analysis and analysis of combined bottom-up processes. Functional classification is due to administrative standardization and the development of bureaucracy since the end of World War II, leading to rationalization and, at the same time, an exponential increase in the complexity of production and use of legal-administrative documents.

This change gives us the foundation to go beyond and think about KO's possible contributions to archival science.

Having provenance as a system premise and functional analysis as a constituent element of the representation system has limitations, especially to users who do not search for information this way. So, is necessary to go beyond the contextual and allows a deepening learn towards the content and the decrease of subjectivity, as we see in Sousa and Araújo Jr. (2013; 2017) when they approach the taxonomies, and in Barros and Gomes (2018) and Gomes et al. (2020) when approaching the ontologies.

A critical factor that makes it possible to apply a KOS in the context of public archives is that most organizational activities are repetitive; they are instances of processes that run frequently.

So, here is where Dahlberg’s triangular concept theory can help us delineate how we name things in archival systems and based on the relationships between terms and concepts in a given juridical-administrative domain, can help construct KOS as a complementary tool to archival classification.

Some authors point out the problem that occurs in relation not only to classification but to a recurrent problem in archival science that can be aided by concept theory and the development of KOS: the naming of classes in research tools and classification schemes. Orr (2005, 111) established that “There is no common rule-based classification model, either in the number of elements or in the levels or the naming of the classes.” Another recurring problem is the lack of deepening theories concerning methodologies of the field; according to Hurley (1993, 11), “The science and methodology of functional
analysis have not yet been written.” Since the mid-1980s, studies have focused on conceptual questions but with few fundamentally methodological reflections. Shepherd and Yeo (2003, 73) write that “Classification schemes are based on an analysis of functions, processes, and activities” ISO15489-1 (2016, 14), which is a records management standard, states: “Classification systems reflect the business of the organization from which they derive and are normally based on an analysis of the organization’s business activities,” and that has been the primary concern in archival theory over time. Nevertheless, when we look to the classification system itself as stated by Foscarini (2006, 191), “the number of classification systems that claim to be function-based, at a deeper glance turns out to be just a mirror of the agency’s internal structure” not reflecting the business functions.

What we argued here is that we can build a process, a flux of activities that help to develop better classification schemes.

The first thing is to construct a policy that established a basis for the whole process, the intellectual work developed by archivists to design the classification scheme, then build a terms-concepts relation with a KOS(i.e., taxonomy) of the institutional domain based on the connection between concepts, terms, and users in a top-down, bottom-up process, then with this cross-reference study the actual construction of the classification with its notations relating concepts, in its triangular Dahlberg idea and the terms they represent in the institutional domain.

With this description seems the job is easy but is not; it is a complex process that, in this manner, has a more balanced approach than the traditional one. We did this across some theoretical-methodological articles such as Barros and Gomes (2018) and Gomes et al. (2020), and the process did work. We urge to build better systems, and this is a possible way to do it.

5.0 Conclusion
In this work, we started from the relation between KOS and concept theory and how these systems and theory can help build better archival classification systems. We argued that it is possible to use KOS and concept theory as part of the traditional classification approach to records. It is a significant issue when we relate how classification is used by non-archivists, which are the ones that produce the records and that have the right to access then. As said in the beginning, this the first time to bring these subjects close in this matter, we will, in the future, stay in this path bring KO closer and closer to archival science theory.

References


Hurley, Chris. 1993 “What, if Anything, Is a Function?” Archives and Manuscripts 21, no. 2: 208-220


Harvesting Knowledge from Cultural Images with Assorted Technologies
The Example of the ChIA Project

Abstract:
In recent years, cross-disciplinary collaboration for increased knowledge extraction from diverse data sources has been at the heart of interdisciplinary research and related fields, such as Digital Humanities. In particular, knowledge extraction and preservation from cultural heritage data has received increased attention. In this paper, we introduce the ChIA project, a cross-disciplinary Digital Humanities project that aims to perform knowledge design, knowledge extraction and organisation by applying semantic as well as Artificial Intelligence (AI) tools on a set of Europeana cultural food images. The collaborative endeavour aims to increase cultural knowledge access and analysis possibilities of images for different user groups and stakeholders, such as content providers or for educational purposes.

1.0 Introduction
Since the recent European Year of Cultural Heritage by the European Commission1, the role of science and technology for the benefit of cultural heritage has played a major role in both physical and digital realms. Not only preserving, but also capturing, organising and making accessible cultural knowledge to different user groups has been a widely addressed topic across disciplines in recent years (cf. Hardman et al. 2009; Behli, Bouras, and Foufou 2018). The ChIA project2 is a current endeavour between the Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH OeAW, AT)3, the Adapt Centre, Dublin City University (IE)4 and the cultural content aggregator Europeana Local - Österreich5. The project, which is at an early stage after inception, focuses on integrating semantic technologies and image analysis to enhance the accessibility of cultural images. Through digitisation, cultural images are being transformed from tangible to digital intangible resources and become accessible via different platforms. This significantly increases their availability to the wider public. The transformation, however, mainly focuses on the conversion of the images into their digital representations with generic metadata (e.g. title, creator, caption), overlooking much of the content of the images themselves. Search and retrieval, thus become restricted only to the generic and associated metadata, failing to explain the relevance of the search results. Often, digital cultural images lack a systematic machine-readable description of the cultural

2 https://chia.acdh.oeaw.ac.at/ [last access: 01.02.2020]
3 https://www.oeaw.ac.at/acdh/ [last access: 20.09.2019]
4 https://www.adapcentre.ie/ [last access: 20.09.2019]
and social aspects, as well as semantic enrichment and a well-defined interlinking of the cultural knowledge embedded. In the ChIA project, we aim to increase the knowledge that can be drawn from such images and enable its presentation in a structured way through the support of different digital tools and methods.

2.0 The method

The ChIA project draws on expertise from an interdisciplinary team, with backgrounds in semantic technologies, artificial intelligence, cultural heritage aggregation and linguistics. In our approach, recent advancements in technology, including deep learning and computer vision (CV), are applied to analyse and capture the content of a selected set of images from Europeana, using pattern matching algorithms to generate new metadata to enable better search and retrieval. Most of these methods, however, focus on automatic object detection typically in the form of predicted concepts and their probability, still with little coverage on the social and cultural aspects of the images. The ChIA system (see Fig.1) aims to combine these different tools and methods that support knowledge extraction and organisation.

Figure 1: Visual representation of the ChIA system. Image @ Yalemisew Abgaz 2019.

Semantic tools enable a unified representation of the entities and foster accurate interpretation; knowledge graphs can be generated by combining all metadata; visual search allows users to search for similar images and a proposed chatbot enables interactive communication exposing the data of Europeana in innovative and new ways. Another aspect that makes the project unique is its specific focus on food-related images (see Figure 2 for an example).
The data, provided by Europeana, has previously been curated and provided to Europeana by cultural organisations, including museums, archives, libraries and galleries. As “food” is a rather flexible concept, we specifically concentrate on food that is edible by humans, and images that present it in a cultural setting, which typically involves persons, locations, objects or a combination thereof. Additionally, “culture” may also encompass depictions of family, societal traditions or customs.

3.0 The aim and scope of the study

The main content objective of ChIA is to develop knowledge design based on Europeana data derived via the Europeana API and explore & analyse cultural content in an experimental setting using AI. In this setting, we aim to enable access and analysis of cultural images by means of widening search capabilities using a combination of semantic technologies and augmented metadata, with interactive tools like chatbots, knowledge graphs and visual analysis. The project more generally functions as a testbed and playground for experimenting with Artificial Intelligence (AI) in a digital cultural context (cf. Schnapp 2014). The new knowledge design models will be beneficial for both scientists and other actor groups, i.e. Europeana content providers or for educational purposes, allowing for much more complex searches than simple metadata-based solutions. Another objective concerns the establishing of an intermediate layer infrastructure with the aim to gain additional metadata from Europeana images, enhance the already existing ones and connect AI services from both industry and open source tools for comparative experimentation. Our approach is novel, in that the combination of proposed tools and the resulting knowledge organisation system, have not been applied to images in the framework of Europeana, a digital cultural heritage collection, before. In addition, the topic of choice, food, is an important aspect of mankind’s tangible and intangible cultural heritage. Making cultural knowledge from images available beyond given metadata enables wider reaching access, but also increased possibilities for various analyses for both humans and machines (Abgaz et al. 2018). Given the time and available resource scope of the project, we aim to test different scenarios and provide test reports.

The resulting knowledge organisation system (KOS), the ChIA system (see also Figure 1), aims to unite the different aspects of semantic and AI technologies, and give users
improved access to as well as improved interaction possibilities with Europeana cultural images depicting food concepts. On the one hand, the system will integrate thesauri with a special focus on food and descriptions of subjects represented in images. Available thesauri will be evaluated and finally reused in a new skosified and open ChIA vocabulary for image description. Thus resource discovery of cultural images that depict food related subjects shall be enabled and improved. The final ChIA food & culture vocabulary will be available in standard SKOS (Simple Knowledge Organization System) format and open for reuse via web services for cataloguing purposes. In line with this, domain-specific relationships between the SKOS concepts will be captured and represented using a separate OWL (Web Ontology Language) ontology. The ChIA food concepts and the ontology will also serve as input to Computer Vision concepts and for further user interaction and data enrichment. On the other hand, the ChIA system also integrates AI results, such as Computer Vision predicted concepts for the selected images analysed. By means of integrating a chatbot, users are also enabled to engage with the images in a more interactive way.

4.0 Conclusion and outlook

Finally, ChIA aims to draw on learnings and make use of larger existing European infrastructures such as DARIAH6 and E-RIHS7. By using cultural images related to edible food from the Europeana database the project can draw on a considerable quantity of cultural heritage data derived from a great variety of cultural content holders (museums, archives, libraries, botanical gardens) across Europe. This will provide a vast starting point for analysis, but also shows the need to precisely target the research for dedicated user groups. In addition, the initial analysis pointed out the importance of the selection process for accumulating the optimal test set of raw data for ChIA purposes. It is envisaged that content analysis will start with investigating metadata richness within the first test sets accompanied by image analysis using computer vision functionality and artificial intelligence.

References


---

6 https://www.dariah.eu/ [last access: 20.09.2019]
7 http://www.e-rihs.eu/ [last access: 20.09.2019]
Knowledge Organization in Historical Information Systems Revisited
Changes in Society, Technology and Expectations 25 Years Later

Abstract:
The concept of Knowledge Organization in Historical Information Systems (HIS) is revisited 25 years after its original presentation at the Third International ISKO Conference of 1994 (Copenhagen, Denmark), before the impact of the World Wide Web revolution was visible. An assessment of these changes and their main implications for HIS is presented, and the concept of HIS is revisited trying to integrate the new challenges. Subsequently, the main KO design offered in the first proposal is redesigned; and, finally, some suggestions are proposed for future research.

1.0 Aims and scope
The aim of this paper is to revisit the topic of Knowledge Organization in Historical Information Systems (HIS) 25 years after its original presentation at Third International ISKO Conference of 1994, Copenhagen (Denmark) (García-Marco 1994), before the impact of the World Wide Web revolution was visible.

It is important to clarify the difference between historical digital libraries, repositories, archives and museums, on the one hand, and HIS, on the other. Both intend to be tools for preserving historical knowledge, but they have strong differences in purpose and methods. The first group stores documents and their metadata to ensure preservation, retrieval and access: libraries, published research; repositories, unpublished research and, in the case of data repositories, analytical information; and digital archives and museums, digital reproductions of original documents and artefacts. HIS pretend to integrate reproductions, data, models, procedures and historiography in the strongest possible way; and must support actual research, not only publishing and distribution. Their aim is creating ‘shared exemplars’ (Kuhn 1970) for scientific historical research.

2.0 Research questions
In particular, three main research questions are addressed in this paper:
— What are the main changes in networking computing brought by the different waves of the web revolution (document-based, social and semantic)?
— Should the concept of HIS be updated to reflect these changes?
— Could the HIS KO model of 1994 be subsequently improved and clarified?

3.0 Methods
The perspective of this paper is theoretical: it tries to incorporate an assessment of the impact of the web technological revolutions into a previous conceptualization of HIS and especially of the role that KOS should have in them (García-Marco 1994). The discussion departs from a consideration of these changes and their main implications for HIS; then, revisits the concept of HIS trying to integrate such implications; and, finally, reconsiders the overall KO design offered in the first proposal.
4.0 Analysis and discussion

As stated before, results are presented in three sections: an analysis of the web revolution impact, a reassessment of the HIS concept and a reconsideration of the overall KO architecture for such systems.

4.1 An assessment of the technological changes and their implications

Since the invention of the web, the triple helix of technology, society and culture has been moving at a higher speed. The web has proved to be an extraordinary platform to fulfill the promises of the information revolution. Around it, a great standardization process has succeeded in the unprecedented digitalization of all alphabets, languages and media, including the mathematical and graphic languages. But, more remarkably, it has brought networking to the centre of the technological efforts and to the heart of the social and cultural life. Therefore, the original HIS vision must be addressed under new lights: the integration of data and documents; the amalgamation of information management and communication; the subsequent explosion in those areas; the incorporation of ontologies to data-and-document organization, processing and communication; and the commoditization of history and culture.

First, thanks to the XML family of standards, modern information management systems have integrated the database and document management functions seamlessly: documents for complex human communication, and data being the raw material for automatic processing. The flow between automatic and human processing is now fluid.

On the part of documents, everything is now susceptible of having digital surrogates, facilitating, through the digitization of original documents, the closest possible contact with the sources that sound historical research requires. Documents of every kind are continuously incorporated to the web. On the part of data, its production has become huge, and it is actually increasing very quickly (e.g., IoT). As a result, the history of our age will be done in a very different way from this of previous ones: historians will work mainly with data as other social scientists, not only with documents, as it was usual in the past. This will finally force a more intense data-gathering approach toward previous ages. The possibilities offered by all this new data have opened new fields of application–open data, big data (Graham, Milligan, and Weingart 2016)–and new scientific disciplines, offering theories, models and methodologies (data science). This new emphasis has been addressed in the LIS field with new concepts and specialties, like data librarians; and data-oriented architects are also needed in the KO field.

Under this abundance of documents and data, there is a great risk for societies and cultures to become buried under such an information overload. Fortunately, the Semantic Web is now providing a full and operational implementation of the information pyramid (knowledge, information, data), and tools for analysing, compacting and making all this information processable. The technology for complex and fully-distributed knowledge systems actually exists, and the expectations for all the stakeholders are very high; but results are still limited. There is a huge and difficult task ahead, and KO experts can reclaim a key role in this quest.

In the particular field of history, the growth of historical databases and information systems of very different types (spatial, heritage-oriented, institution-oriented…) has in fact made the problem of knowledge organization even more acute, as each system
provides a divergent KOS. On the other hand, History as a discipline is becoming more complex and new approaches are needed from the KO arena (Gnoli 2014).

Challenges also come from the new social relevance of History, which is becoming a phenomenon for the masses. History, far or recent, has become increasingly the raw material of mass-communication and entertainment products and services (TV series, films, tourism, fashion…). This is both an opportunity and a challenge for historical sciences, professionals and the general public, but, on the contrary, as we can appreciate daily, there are cons. History and fiction get dangerously mixed according to the agendas of producers, distributors and audiences, which has very serious implications for evidence-based identity construction and preservation. Evidence interlinking and data comparison can provide light and truth to this confusing environment, so that each person can think and decide on its position informedly. KOS also seem central to this interlinking project among different sources and media.

4.2 Must the HIS concept change?

Then, a HIS was defined as (García-Marco 1994)

[…] an automated system integrating a set of databases and formal procedures, designed and maintained to store, treat and retrieve historical information. They must be able to store sources —both references and reproductions—, bibliographical references and research work; should they be textual, graphic or procedural. The stress is put in the interface among the different types of information. The HIS is considered to be an open evolutionary system, growing towards an ever-closer integration of the data. Finally, the HIS must be considered as a part of a scientific network of research and custodial centres, with which it must interchange data and knowledge, and therefore pursue co-operative normalization policies.

It is clear that in 1994 the focus was still on stand-alone systems, centralized or synchronized, and with distributed access; though the interoperation among research and custodial centres was foreseen. This concept was useful because it contemplated the integration of documents (sources and historiography), data, procedures and models. But any reassessment in 2020 requires a distributed networking perspective, where different HIS cooperate, compete or simply coexist with a potential for resource sharing.

4.3 Must the overall KO vision for HIS change?

Among other reasons, history is a very interesting knowledge domain because at the same time it is a ‘natural’, encyclopaedic (it encompasses all the reality) and objective domain (all that happened and was in the past), and a discipline with many possible perspectives and subdisciplines (the study of history).

History becomes a natural concept as soon as memory develops: all events and entities will be history; all information systems will become historical sources; and, certainly, there is room for a historical subdiscipline in every scientific domain. On the other hand, history has become a highly academic field, and interdisciplinarity is now at the heart of historical specialization. Regarding the problem of interdisciplinarity, the matrix analysis proposed by Dahlberg (2008) is strongly relevant, though transdisciplinary problems are provoking the emergence of a new layer (e.g., urban studies).

On the problem of the connection between the two layers, the ideas of Claudio Gnoli on the classification of phenomena and the use of facets to connect them with the layer of refection and research, e.g., disciplines, are very relevant (Gnoli 2016; 2017a; 2017b; 2018). This vision was previously expressed in The León Manifesto (Gnoli and Szostak,
2007), which we share completely, though we take a neutral stage on the precedence of phenomena over disciplines. This is a very interesting problem in history, where the object-subject relation is especially evident and even intense, and, where, except for the common realities still shared by old and current cultures, the identification of phenomena is, many times, an act of creative interpretation. Apart from this, most historians develop their work inside disciplines, and only some of them are theoretically implied in the study of phenomena in genuine interdisciplinary ways.

Table 1. Main layers and facets to consider in HIS: an update

<table>
<thead>
<tr>
<th>Transversal facets</th>
<th>History</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) Persons (historical)</td>
<td></td>
</tr>
<tr>
<td>b) Spatial context</td>
<td></td>
</tr>
<tr>
<td>i) Temporal context (objective)</td>
<td></td>
</tr>
<tr>
<td>g) Other live beings</td>
<td></td>
</tr>
<tr>
<td>c) Objects and artefacts</td>
<td></td>
</tr>
<tr>
<td>e) Actions and processes</td>
<td></td>
</tr>
<tr>
<td>*) In-text subject tags</td>
<td></td>
</tr>
<tr>
<td>Historiography</td>
<td></td>
</tr>
<tr>
<td>b) Persons (historiographical)</td>
<td></td>
</tr>
<tr>
<td>i) Temporal context (periodizations)</td>
<td></td>
</tr>
<tr>
<td>g) Disciplinary subdivisions (Historical disciplines)</td>
<td></td>
</tr>
<tr>
<td>j) Theoretical systems and paradigms</td>
<td></td>
</tr>
<tr>
<td>k) Methodologies</td>
<td></td>
</tr>
<tr>
<td>*) Tools</td>
<td></td>
</tr>
<tr>
<td>f) Great themes or ‘social institutions’: society, economy, religion, culture, arts, policy, military and war, science, technology, etc.</td>
<td></td>
</tr>
<tr>
<td>[d) Specific scientific categories and constructs applied historical objects and singulars]</td>
<td></td>
</tr>
<tr>
<td>d1) Actions and processes</td>
<td></td>
</tr>
<tr>
<td>d2) Entities</td>
<td></td>
</tr>
<tr>
<td>Sources</td>
<td></td>
</tr>
<tr>
<td>Users</td>
<td></td>
</tr>
</tbody>
</table>
| *) Purpose (research, teaching, political…)

At this stage, we have differentiated clearly historical phenomena (history) and historiography as two distinct levels, and added a few general facets to them (denoted with an asterisk) (Table 1). We also think that any KOS of HIS should incorporate historical sources as a distinct layer, though there is some overlapping with the historiographical layer (see italics). This layer is not developed, as it gets into the difficult arena of functional requirements, their subsequent cataloguing models (LRM-RDA, EGAD, CIDOC-CRM…) and their connection to subject indexing and classification. Moreover, users should be incorporated as a domain, though this is also partially implied in the historiographical domain. With this new layer, the four elements of the basic ontology for History as a discipline would be available: history (phenomena) is told (historiography) to users from sources (hopefully).
5.0 Conclusion

The web revolution has inserted HIS into more complex information ecologies: HIS will not evolve as independent units but as networking systems.

Thanks to the WWW and the work of so many persons and institutions, great networking digital library, archive and museum systems have appeared (LC, Europeana, etc.) that are making this vision a reality in many of its layers. The digital humanities movement is also promoting huge advances in History Informatics research and applications (e.g., Meyer, Schering, and Schmitt 2014; Graham, Milligan, and Weingart 2016). However, the integrative aspects that were central to the vision of HIS have a long way to be fully developed. Here, KO expertise could be decisive after many years devoted to the problems of interdisciplinarity and KOS interoperability.

But this task requires alliances and cooperation outside the traditional KO field. The efforts in the design and development of KOS for the historical sciences must be integrated with the recent great steps forward brought by the conceptual models for the description of resources in libraries, museums and archives; and also, with the advances in the semantic web and the big data developments.

Acknowledgments

This research work was carried within the CSO2015-65448-R (MINECO/FEDER) project. We are also grateful to the anonymous referees for their invaluable suggestions.

References


Representing Entities and Characteristics of Iranian Performing Arts Based on IFLA Library Reference Model (IFLA-LRM)

Abstract:
IFLA-LRM is the result of consolidating three separate FR family models. The purpose of this paper was to represent Iranian performing arts characteristics based on IFLA-LRM. This paper is an evaluative survey. The research population comprised of 1368 performing arts records in NLAI. A checklist was used to gather the required data. The results showed that there is a good adaptation between performing arts resources with manifestation, item, nomen, agent, time span, and place entities of LRM. Also, the results showed that there were not accurate links from the work itself to its expressions. Some gaps in identifying “works” of performing arts and their characteristics through LRM attributes, and relationships were observed. Although LRM has sufficient entities, attributes and relationships, Application of LRM to Iranian performing arts remains problematic. These problems may be explained by problems of considering performance as a work and library software problems in representing essential fields. Some amendments in work and expression levels and their attributes in LRM will be needed to better describe performing arts.

1.0 Introduction
In last years, according to the new technologies such as semantic web and linked data, the bibliographic universe has seen great changes in users’ needs and how Knowledge Organization Systems (KOS) model bibliographic data in information systems. Therefore, in the 1990s, IFLA and other organizations started to work on conceptual models. Functional Requirements for Bibliographic Records (FRBR) was the first conceptual model which was endorsed in 1998. FRBR is an entity-relationship model. Then, FRAD and FRSAD were developed to represent authority data and subject authority data in the bibliographic universe. Since endorsing FRSAD in 2010, it turned out that utilizing a unified model is easier than implementing three different models. Therefore, in 2010, the FRBR Review Group started to work on a consolidated model. Finally, in August 2017, three models superseded by IFLA Library Reference Model (Riva, Le Boeuf, and Žumer 2017). IFLA-LRM was developed to resolve inconsistencies between three separate models. This model covers all aspects of bibliographic and authority information (Riva and Žumer 2018). LRM facilitates the harmonization among the different cataloging standards (Dunsire et al. 2018). This model seeks to provide a conceptual framework for describing bibliographic entities. LRM focuses on user needs to create online catalogs that better support user information seeking. IFLA LRM has 11 main entities, 37 attributes, and 36 relationships.

2.0 Problem statement
Documenting performing arts is a rapidly developing field, and has undergone a radical shift in the digital age (Lee 2018). It is impossible to hold live performances in libraries. Therefore, these materials are rarely considered in cataloging standards (Miller and Le Boeuf 2005). Whereas, Libraries hold materials of performance such as theater, music performances, and so on through their recordings or other forms of materials.
These resources are related to the library, archive, and museum communities. Therefore, it is necessary to pay more attention to their characteristics and how they are represented in the information systems. By considering the importance of materials related to performing arts, it will be useful to analyze how Iranian performing arts fit into conceptual models. The FRBR and LRM models are important for bibliographic description and access. Therefore, it will be useful to understand how they fit into materials related to performing arts.

3.0 Literature review

Literature concerning modelling performing arts as well as implementing conceptual models was reviewed. Some researches have been conducted in these fields. Riva, Le Boeuf, and Žumer (2017), Riva and Žumer (2017), Oliver (2018), Dunsire et al. (2018) are of great importance for understanding the concepts of IFLA-LRM. The studies by Taniguchi (2018) and Žumer, Aalberg and O’Neill (2019) showed how conceptual models fit into different resources. Likewise, some researches have been conducted in completing conceptual models. For instance, In the case of performing arts, Miller and Le Boeuf (2005), Doerr, Le Boeuf, and Bekiari (2008), Lee (2018) studied opportunities of conceptual models in modelling performing arts. For instance, Miller and Le Boeuf (2005) discussed the performing arts and their entities in the FRBR model. In other research Lee (2018) analyzed Miller and Le Boeuf’s (2005) model and showed issues of their approach. Also, Doerr, Le Boeuf, and Bekiari (2008) analyzed how conceptual models could be used to describe performing arts. They proposed a detailed model based on CIDOC CRM and FRBRoo model to organize performing arts materials. FRBRoo is the result of harmonization between CIDOC CRM and FRBR and utilizes object-oriented formalism. Literature shows that Most projects examined conceptual models to describe musical works and performing arts are absent in these examinations. As performing arts are related to the culture of a country, it will be useful to analyze regional performing arts and their characteristics based on conceptual models.

4.0 The aim and scope of the study

This study aims to represent entities and characteristics of Iranian performing arts based on IFLA-LRM. To achieve this goal, 3 main questions have been defined:

• What are the main characteristics and entities of Iranian performing arts?
• How can we analyze the opportunities of IFLA-LRM in describing Iranian performing arts?
• What are the specific characteristics of Iranian performing arts which do not already exist in IFLA-LRM?

5.0 Methodology

This paper is an evaluative survey. The research population was comprised of 1368 performing arts records in the RASA software. RASA is important library software in IRAN and produced by Pars Azarakhsh Company. RASA has been tailored especially for the National Library of Iran. This software has been based on IRANMARC which is based on UNIMARC in turn. In this study, a checklist was used to gather and analyze the required data. This checklist was based on the entities and characteristics of Iranian performing arts and entities of the Library Reference Model (LRM). Mapping tables of
LRM to other conceptual models were used to be sure about the accuracy of gathered data. By analyzing the checklist, the degree of adaptation between Iranian performing arts with IFLA-LRM has been analyzed. Also, in this paper, some implementation tips about Iranian performing arts have been suggested to complete IFLA-LRM.

6.0 Results

To gather the required data performing arts records in RASA software were studied. Table 1. shows the frequency of Iranian performing arts in this library system.

<table>
<thead>
<tr>
<th>Iranian Performing Arts resources</th>
<th>Types of resources</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Document</td>
<td>Audio visual</td>
</tr>
<tr>
<td>Opera</td>
<td>347</td>
<td>3</td>
</tr>
<tr>
<td>Orchestra</td>
<td>69</td>
<td>135</td>
</tr>
<tr>
<td>Ballet dancing</td>
<td>217</td>
<td>12</td>
</tr>
<tr>
<td>Folklore Dance</td>
<td>48</td>
<td>44</td>
</tr>
<tr>
<td>Siah bazi</td>
<td>7</td>
<td>57</td>
</tr>
<tr>
<td>Naghali</td>
<td>20</td>
<td>14</td>
</tr>
<tr>
<td>Theatre</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>Pardeh khani</td>
<td>-</td>
<td>6</td>
</tr>
<tr>
<td>Ruhowsi</td>
<td>-</td>
<td>22</td>
</tr>
<tr>
<td>Kheyme shab bazi</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Taziye khani</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Mareke giri</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>Aroosak gardani</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Chavosh khani</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Ghashogh zani</td>
<td>-</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 1 shows the variety of Iranian performing arts in the RASA software. There were 1368 records related to Iranian performing arts. The most frequent Iranian performing art was Opera. Orchestra and Ballet dancing were in the second and third positions respectively. The blow table shows entities of Iranian performing arts resources.

The main fields to identify entities of Iranian performing arts in association with LRM were titles, uniform titles, statement of responsibility, physical description, publication information, and added entries. “Item” entities were found based on keeping place and registration number of resources. Audiovisual resources consist of different forms such as compact Discs, images, cards, and so on. Their entities identified according to their title, agents, descriptors, compilers, photographers, directors, and so on. As shown in table 2. There is a good adaptation between performing arts resources with manifestation, item, nomen, agent, time span and place entities of LRM. There were not accurate links from the work itself to its expressions. Only in the textual plays, the link between expressions and their works observed through uniform titles. LRM describes the form of work through the category attribute. Also, carrier type of resources is declared in the category of carrier attribute of manifestation. Performance is limited to a certain time and place. Therefore, performance is contained within conceptual models through its
recordings. In the conceptual models, a recording is considered as an expression of work.

Table 2. Entities of Iranian performing arts resources

<table>
<thead>
<tr>
<th>Type</th>
<th>Freq</th>
<th>Work</th>
<th>Expression</th>
<th>Manifestation</th>
<th>Item</th>
<th>Nomen</th>
<th>Agent</th>
<th>Person</th>
<th>Collective agent</th>
<th>Time-span</th>
<th>Place</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents</td>
<td></td>
<td>741</td>
<td>116</td>
<td>741</td>
<td>711</td>
<td>741</td>
<td>650</td>
<td>746</td>
<td>10</td>
<td>-</td>
<td>148</td>
</tr>
<tr>
<td>Audio visuals</td>
<td>-</td>
<td>311</td>
<td>56</td>
<td>300</td>
<td>280</td>
<td>287</td>
<td>215</td>
<td>300</td>
<td>200</td>
<td>-</td>
<td>30</td>
</tr>
<tr>
<td>Books</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full text</td>
<td>132</td>
<td>31</td>
<td>132</td>
<td>14</td>
<td>89</td>
<td>200</td>
<td>132</td>
<td>112</td>
<td>14</td>
<td>132</td>
<td>132</td>
</tr>
<tr>
<td>Proceeding</td>
<td>20</td>
<td>2</td>
<td>20</td>
<td>2</td>
<td>14</td>
<td>19</td>
<td>16</td>
<td>19</td>
<td>3</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Adaptation</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>-</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>Summery</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>The play</td>
<td>38</td>
<td>12</td>
<td>38</td>
<td>10</td>
<td>35</td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>-</td>
<td>36</td>
<td>37</td>
</tr>
<tr>
<td>Stories</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>-</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Poem</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>-</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>-</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Theses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>71</td>
<td>28</td>
<td>71</td>
<td>69</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>183</td>
<td>-</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>Oral history</td>
<td>-</td>
<td>41</td>
<td>39</td>
<td>41</td>
<td>35</td>
<td>41</td>
<td>120</td>
<td>41</td>
<td>-</td>
<td>-</td>
<td>41</td>
</tr>
<tr>
<td>Serial</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

7.0 Conclusion

The main aim of this paper was to identify the degree of adaptation between IFLA-LRM and Iranian performing arts. The results of the study revealed that there is a good adaptation between performing arts resources with item, nomen, agent, time span, and place entities of LRM. This low compatibility may be explained by problems of considering performance as a work and library software problems in representing essential fields. There are some gaps in describing “works” of performing arts. Performing work itself is still unclear in LRM. Catalogers deal with manifestations and items in their institutions. There were not accurate links from the work entity to its expressions. Making links between a “performance work” and its various mediums is important. It is
essential to pay more attention to a recording of performing arts and their various forms of carriers. There must be a link from recording to the performance event. Also, a clear distinction between performer and creator is needed. There are some gaps in the “performance work” of IFLA-LRM. For instance, the medium of performance attributes in LRM has pertained to musical works. This attribute is not fully accounted for various forms of art. It is suggested to expand this attribute. Whereas, further modifications in work and expression level is required to implement LRM to performing arts.

8.0 Suggestions for further studies

It is suggested to conduct comparative researches about opportunities of FRBRoo and IFLA-LRM in describing performing arts.

It is suggested to implement IFLA-LRM in other kinds of resources and identify its adaptation degree with other resources.

References


An Ontology for Conceptual Analysis of Signature Pedagogies

Abstract

Signature pedagogy is a multifaceted concept that has been found useful by education researchers for analyzing and characterizing pedagogical methods for particular professions. A signature pedagogy must impart not just the formal disciplinary knowledge but also the cognitive, behavioral and attitudinal norms of the profession. The pedagogy is characterized by four dimensions of surface structure, deep structure, implicit structure and shadow structure. The abstract and multifaceted nature of the concept affords researchers latitude in construing and applying the concept in analyzing and describing pedagogies, making it difficult to compare pedagogies across professions and across academic papers. Based on a conceptual analysis of twenty-four journal papers describing signature pedagogies of various professions, the paper proposes and outlines a Research in Signature Pedagogy (ReSiPe) ontology.

1.0 Introduction

Signature pedagogies of the professions is a concept introduced by Shulman (2005a; 2005b) to refer to types of teaching for educating students into particular professions, by imparting not just formal disciplinary knowledge but also ways of thinking (habits of the mind), ways of performing (habits of the hand), and ethical values (habits of the heart) that are normative in the profession. A signature pedagogy is characterized by four dimensions of surface structure, deep structure, implicit structure and shadow structure. The concept has captured the imagination of education researchers as it points to subtle cognitive, behavioral and attitudinal attributes that are expected in a competent member of the profession, and thus should be imparted using appropriate teaching methods.

Signature pedagogy is an abstract and multifaceted concept. Shulman did not define the concept very precisely, and researchers have applied the concept differently according to their analytical purpose and the characteristics of the profession they are analyzing. However, the abstractness and ambiguity of the concept make it difficult to compare signature pedagogies across professions and across academic papers. It is also difficult to apply the concept in designing professional education programs.

We carried out a conceptual analysis of the concept as used in twenty-four journal papers analyzing signature pedagogies of twenty professions/disciplines. The purpose of the conceptual analysis was to clarify what a signature pedagogy is, its subcategories and parent categories, its component parts, and its attributes and aspects. Based on the conceptual analysis, we are developing an ontology that researchers can use to analyze, evaluate and compare signature pedagogies described in academic papers. Shulman (2005a, 58) noted that “the comparative study of signature pedagogies across professions can offer alternative approaches for improving professional education ...”

2.0 Method

An initial coding scheme was constructed based on a close reading of a set of Shulman’s (2005a; 2005b; 2005c; 2006; 2008) seminal papers and four journal articles describing four signature pedagogies. The coding scheme was then used to analyze ten
more papers. A Microsoft Excel spreadsheet was used to analyze each article. The analysis was carried in several iterations: 1) Reading through each article to extract passages that correspond to categories in the coding scheme; 2) Identifying significant concepts in the passages tagged with the coding categories; 3) Analyzing issues relating to each group of categories. Based on the results of the conceptual analysis, an initial Research in Signature Pedagogy (ReSiPe) ontology was constructed and used as an analytical framework to analyze ten more papers to evaluate its utility and completeness. An outline of the ontology is given in Table 1.

3.0 Conceptual analysis of descriptions of signature pedagogies

3.1 Concept of signature pedagogy

Shulman conceived of signature pedagogy as being inextricably identified with a particular profession. However, our analysis of the sample papers indicate that it is difficult to pinpoint the essential characteristic of a pedagogy that is tied to the profession (with the obvious exception of disciplinary content knowledge). Descriptions of signature pedagogies involve abstractions that can conceivably be applied to other disciplines and professions. Authors have relaxed the definition of the concept to apply to subject disciplines (e.g., political science, history and humanities). We propose a broader concept, distinctive pedagogy, to abstract the characteristics of a family of related signature pedagogies across professions. Some authors have narrowed the scope of the concept to particular schools or locales, rather than to the whole profession. To represent innovations in pedagogy, we adopt the idea of emerging or aspiring signature pedagogy (Chick et al. 2012; Horn 2013) to refer to new pedagogical methods developed by particular schools. It will also be useful to identify pedagogies of individual gifted teachers. A profession will also exhibit different characteristics in different countries, indicating that a signature pedagogy should have a location attribute. We propose therefore a class hierarchy with Distinctive pedagogy as the top-level class (see Table 1).

Shulman argued that a signature pedagogy should be pervasive in the sense of being adopted in most professional schools, be routine throughout an education program, have a high level of engagement among students and instructor, resulting in a degree of uncertainty within each instructional session. In the ontology, these are represented as attributes of signature pedagogy.

3.2 Pedagogical structure: surface, deep, implicit and shadow structure

Descriptions of signature pedagogies in the sample papers typically include substantial details of the surface structure, including the component teaching methods (e.g., fieldwork, tutorial and rounds) and characteristics relating to use of space and time, sequence of activities involved, participant roles (e.g., facilitator, coach, expert), and artifacts and technologies employed. The sample papers also discuss the strengths and weaknesses of the surface characteristics, as well as relate some characteristics to professional competencies that they help to impart. Some papers discuss the possible underlying explanation/mechanism of how surface features help students to acquire particular competencies. The papers highlight the distinctive characteristics of a pedagogy that account for its effectiveness and that distinguish it from other pedagogies.

Characteristics of the deep structure (underlying assumptions about how to impart professional competencies) and implicit structure (beliefs about professional attitudes
and values) are abstract and theoretical concepts. Deep structure is often specified as principles relating to pedagogy execution, student and learning, and teacher and teaching. They are often related to habits of the mind. In contrast, implicit structure is often related to habits of the heart. The shadow structure reflects the characteristics that are lacking in the pedagogy. However, what is missing may be positive (a strength) or negative (a weakness).

3.3 Types of professional competencies

According to Shulman (2005a), a signature pedagogy seeks to impart four types of professional competencies: 1) habits of the mind (i.e. ways of knowing and thinking); 2) habits of the hand (i.e. ways of performing and doing things); 3) habits of the heart (i.e. ethics and attitudes); and 4) pedagogical content knowledge.

In the sample papers, habits of the mind tend to be specified as generic (transferable) skills (e.g., critical thinking), research skills, and development skills. The sample papers are generally not successful in characterizing how transferable skills take different forms in different professions. The sample papers are also weak in specifying habits of the hand. It is difficult to disentangle habit of the hand from habit of the mind, as educational programs focus on the professional thinking and judgement that inform practice.

Habit of the heart is often taken to refer to ethics and values, but have been expanded to include attitude, disposition, and identity/being. Attitude appears to represent the professional’s stand or relation towards objects of significance for the profession, leading to a particular behavioral disposition. Identity and being suggest professional self-awareness and self-image, including a sense of the professional’s place and role in society and in the world.

A signature pedagogy must of course impart the formal disciplinary knowledge, but this must be tempered with an understanding of learning and teaching, as well as practical experience of applying the knowledge to form pedagogical content knowledge.

4.0 Utility of the ontology

The initial ReSiPe ontology was encoded in Turtle format using the TopBraid Composer ontology editor, and used to analyze an additional ten journal papers. Preliminary results indicate that the coding can readily reveal which aspects of a signature pedagogy are lacking in descriptive details. However, to adequately compare signature pedagogies, the class hierarchies for various concepts need to be fleshed out further. We believe it is feasible to compile comprehensive class hierarchies for surface structure features, i.e. Teaching method, Activity, Participant role, and Artifact and Technology used. More difficult will be to identify the different habits of the mind, habits of the heart, and teaching/learning principles (of the deep structure). More work is also needed to model the underlying explanations of how surface features help to impart a competency.

A graphical visualization web application based on the ontology is being developed to support educational researchers to systematically analyze and visualize a signature pedagogy, and compare its characteristics with those of the twenty-four signature pedagogies that we have analyzed and coded as ontology instances. The tool is also expected to help teachers to understand and compare pedagogical methods.
Table 1. Research in Signature Pedagogy (ReSiPe) ontology: An outline

*Note: relations are indicated in italics.*

### Distinctive pedagogy
- **Signature pedagogy**
  - Signature pedagogy of a profession (*educatesFor* Profession)
  - Signature pedagogy of a discipline (*educatesFor* Discipline)
  - Signature pedagogy of a school (*associatedWith* Educational institution)
  - Signature pedagogy of a teacher (*associatedWith* Teacher)
- **location**: Geographic location or Educational institution
- **structure**: Pedagogical structure
- **imparts competency**: Professional competency
- **engagement characteristic**:
- **uncertainty characteristic**:
- **pervasiveness characteristic**:
- **routineness characteristic**:
- **distinctive characteristic**: Pedagogical structure or Competency

### Pedagogical structure
- **Surface structure**
  - Teaching method: Teaching method
  - space characteristic:
  - temporal characteristic:
  - activity: Activity
  - participant role: Participant role
  - content:
  - artifact: Artifact
  - technology: Technology
  - **imparts competency**: Professional competency
- **Deep structure** (with relation *imparts competency*: Professional competency)
- **Implicit structure** (*imparts competency*: Habit of heart)
- **Shadow structure** (*lacking competency*: Professional competency; *lacking structure*: Pedagogical structure)
  - Positive shadow
  - Negative shadow

### Professional competency
- **Habit of the mind**
  - Transferable skill
  - Research skill
  - Development skill
- **Habit of the hand**
- **Habit of the heart**
  - Ethics/values
  - Attitude
  - Identity/being
  - Disposition
- **Content knowledge**
  - Academic and research knowledge
  - Pedagogical content knowledge
  - Professional knowledge
  - Classroom knowledge

© Ergon - ein Verlag in der Nomos Verlagsgesellschaft
References


Classifying Perspectives
Expressing Levels of Knowing in the Integrative Levels Classification

Abstract:
Recent classification research calls for multi-perspective knowledge organization systems that support users to identify and interrelate different authorial perspectives on the same subject matter (Gnoli 2011; Kaipainen and Hautamäki 2011; Szostak 2014). In opposition to inductively derived modes of perspectives or viewpoints, Kleineberg (2013, 2014, 2017, 2018) proposes the organizing principle of integrative levels of knowing that takes recourse to already existing models of cognitive development deduced from rational reconstructions of learning processes that follow an invariant sequence of thought styles. This contribution offers some preliminary thoughts on a potential application of such levels of knowing to the Integrative Levels Classification (ILC) under consideration of its core elements like dimensions, levels, and facets.

1.0 Introduction
Documents concerned with the same subject matter can be created by authors taking very different perspectives. Users of information systems should be supported to identify and interrelate such authorial perspectives in order to make relevance decisions or to explore the so-called epistemological dimension of human knowledge (Kleineberg 2013). A particular challenge for knowledge organization systems that seek to incorporate this dimension is the sheer complexity of potential perspectives that can be defined in very different terms and at very different levels of specificity.

For example, Swift, Winn, and Bramer (1978, 186) propose a “multi-modal approach” to the indexing of documents that is not only concerned with the aboutness of research literature but also with underlying theoretical orientations and methods of research (cp. Biagetti 2006; Szostak 2014). The theoretical orientation can be specified, for instance, in terms of approach (e.g., structural-functionalism) and discipleship (e.g., Durkheimian).

In a similar way, Austin’s (1984, 168) PRECIS thesaurus entails an operator called “viewpoint-as-form” that distinguishes between viewpoint, perspective, and aspect. While viewpoint refers to a class of people (e.g., Christian viewpoint, Trade Union viewpoint), perspective corresponds to a discipline (e.g., sociological perspective, philosophical perspective), whereas aspect means a particular focus on a given knowledge subject in special studies (e.g., economic aspect, social aspect).

A further influential distinction is introduced by Langridge (1989) between topic and form of knowledge. While the topic is related to perceived phenomena, the form of knowledge describes the way in which these phenomena are perceived. Langridge (1989, 45) identifies a broad range of such “formal characteristics” that include the perspective from which a document is created in terms of religious or ideological world view (e.g., Christianity, Marxism), philosophical viewpoint (e.g., rationalism, empiricism, pragmatism, realism, idealism, humanism), school of thought (e.g., behaviorism and Gestalt school in psychology), epoch (e.g., ancient, medieval, modern), culture (e.g., Western, Chinese, Indian), intellectual level (e.g., elementary, advanced), etc.
or rhetoric form (e.g., description, analysis, interpretation, narrative, prescription, evaluation).

But as Taylor and Joudrey (2017, 463) in their textbook The Organization of Information conclude, document characteristics like point of view or perspective are “rarely if ever translated into controlled vocabulary terms or classification notations; instead, it may be addressed in summary statements or abstracts.” In the following, it will be argued that integrative levels of knowing offer a promising tool for a systematic, comprehensive, and relatively stable organization of authorial perspectives, and it will be demonstrated how this, as controlled vocabulary, could be adopted by the Integrative Levels Classification.

2.0 Levels of knowing

The above-mentioned modes of perspective (i.e., theoretical approach, discipleship, class-related or philosophical viewpoint, disciplinary perspective, religious or ideological worldview, school of thought, epoch, culture, or rhetoric form) have in common that their underlying typologies need to be inductively derived from the literature and will hardly arrive at the status of exhaustive classifications. This is because of the dynamic and overlapping nature of these kinds of authorial perspectives (problem of collocation or class-building) and the missing hierarchical relations between different types within a given mode of perspective (problem of subsumption or subclass-building).

The advantage of integrative levels of knowing as a mode of perspective is that there already exist well-elaborated models of cognitive development that offer solutions for both problems. The idea is best-known from, but not limited to, the Piagetian tradition of developmental structuralism (Habermas 1979; Barnes 2000). Accordingly, developmental stages or levels of knowing present qualitatively distinct cognitive structures or styles of thought along an invariant sequence of increasingly differentiated and integrated structures (Kleineberg 2014). In other words, models of integrative levels of knowing present well-defined thought styles or perspectives (class-building) that are arranged as hierarchical integrations (subclass-building). This provides a useful organizing principle to reduce the diversity of perspectives to a manageable amount of cognitive structures or thought styles, as Michael Barnes (2000, 45; emphases his) suggests:

“Piaget's description of stages helps to categorize thought styles more clearly. His theory also helps to recognize the particular sequence in which stages of thought appear in cultural history. In cultures as in individuals, the easier modes of thought appear first and continue to be used even when more difficult modes of thought are added.”

In fields like psychology, sociology, anthropology, and cultural studies, models of integrative levels of knowing are rationally reconstructed for individual and collective learning processes that cover a broad range of domain-specific cognitive competences, such as mathematico-logical reasoning, physical thinking, aesthetic judgment, ego identity, self-understanding, social cognition, moral consciousness, and religious thought (for an overview of several dozen models see Wilber 2000 and Kleineberg 2017). Empirical findings suggest that these kinds of basic cognitive structures or thought styles are cross-culturally and across historical times valid (Habermas 1979; Barnes 2000). This means that these cognitive-developmental models present comprehensive and relatively stable (rational reconstructions are still fallible)
classifications of perspectives that offer useful tools for multi-perspective knowledge organization systems (cp. Esbjörn-Hargens and Zimmerman 2009; Kleineberg 2018).

3.0 Integrative Levels Classification

The experimental Integrative Levels Classification is an ongoing international research project under the lead of Claudio Gnoli and presents one of the most ambitious attempts to express multiple viewpoints or authorial perspectives within a comprehensive, interdisciplinary, and phenomenon-based knowledge organization system (Gnoli 2020). The core elements of ILC are dimensions, levels, and facets.

Dimensions define different kinds of metadata that describe phenomena, perspectives, documents, collections, information needs, and people (Gnoli 2016). The ontological dimension of phenomena functions as the basis for the classification scheme and can be related to the other dimensions by using the analytico-synthetic technique of freely combinable facets. For the incorporation of the organizing principle of integrative levels of knowing, the epistemological dimension of perspectives is of most importance and in the focus of this paper. However, levels of knowing can also be treated as phenomena, for example, in regard to books on cognitive development; or they may be applied to the sociological dimension of people, for example, in regard to user groups like children (cp. Beak 2014).

Levels define the main classes of the classification scheme reflecting phenomena along a sequence of increasing organization and evolutionary order, inspired by the theory of integrative levels (Gnoli 2017a). The main classes are indicated by lower-case letters from a to z (e.g., a forms, e atoms, f molecules, l bacteria, m organisms, p consciousness, s communities, w artifacts, y knowledge) and are divided into several subclasses or types as genus-species relations by adding further lower-case letters to the notation (e.g., mq animals, mqvo birds).

Facets define particular properties or relationships of a given phenomenon in terms of fundamental categories or subcategories indicated by digits (Gnoli 2017b). The recent second edition ILC2 includes the following fundamental categories: 0 perspective, 1 time, 2 place, 3 agent, 4 disorder, 5 transformation, 6 property, 7 part, 8 form, and 9 kind. The perspective facet 0 is intended to embrace the whole epistemological dimension in terms of viewpoint, aspect, bias, discipline, theory, method, domain of discourse, activity field, locus and epoch of knowledge, and cultural context (Gnoli 2016, 2017b). But its subclasses, indicated by a combination of several digits (e.g., 05 activity field/sphere/domain, 053 discourse community, 059 culture/civilization), seem to express only those modes of authorial perspective that are described above as inductively derived and less systematic.

This paper proposes to incorporate the more systematic integrative levels of knowing as a further subcategory of the perspective facet. Since the different values or foci of a given facet are supposed to be taken from any other place of the ILC’s scheme, there are two potential classification notations to begin with, one for the individual dimension and one for the collective dimension of cognitive development. The individual dimension can be related to main class p consciousness and its subclass po cognition,

1 http://www.iskoi.org/ilc/
2 http://www.iskoi.org/ilc/2/ilc.php
where already subclasses exist like por reason or pou moral consciousness. Likewise, the collective dimension can be related to main class y knowledge and its subclass yas style of thought, where already subclasses exist like yasm magic thought, yass scientific thought, or yasw world views.

For both dimensions, it would be easily possible to expand the ILC’s scheme with subclasses that borrow their terminology from existing cognitive-developmental models with their well-defined and technically named levels of knowing (e.g., Kohlberg’s moral consciousness: preconventional—conventional—postconventional; Habermas’s world views: magical-animistic—mythological—rationalized—reflexive).

Classifying authorial perspectives could then look like this example: John Rawls’s A Theory of Justice presents a document on the subject matter or phenomenon of justice theory, created by an author taking a perspective of mature postconventional moral consciousness according to Kohlberg’s stage or level 6 called “universal ethical principle orientation” in his model of individual moral development (cp. Kleineberg 2018). In ILC, this could be expressed by a combination of the phenomenon class yisral justice theory (knowledge > disciplines > empirical sciences > sociology > justice theory) and the perspective facet 0pouv universal ethical principle orientation (consciousness > cognition > moral consciousness > universal ethical principle orientation), resulting in the classification notation: yisral0pouv.

4.0 Conclusion

After these preliminary thoughts, some important open questions remain. First and foremost, how could integrative levels of knowing be defined as a special subcategory and placed within the perspective facets? Second, should all domain-specific models of individual (or collective) cognitive-development subsumed under the same class, for example, por reason (or yas style of thought)? And finally, which of the existing models of cognitive development should be adopted as controlled vocabularies at all?

References


Linking, Mapping, Matching, and Change
Contemporary Use of Ranganathan’s Three Planes of Work in Classification Activity

Abstract:
Scholars have identified interoperability issues in mapping metadata in a linked data environment (Zeng 2019). This study builds on previous research and proposes a creative use of Ranganathan’s (1989) three planes of work in classification activity. By extending the application of the three planes of work to the linked data environment, we can use this conceptual model as an analytical tool to highlight particular mapping challenges. This paper uses three cases to show how discrepancies between the idea plane, verbal plane, and notational plane may cause mapping issues. Further, we can see that mapping issues are not limited to differences between metadata standards. The three planes of work can highlight mapping issues that are caused by changes at different planes of the same metadata. The challenges presented in this study complement the known mapping issues, and contribute to the discussion of interoperability in linking, mapping, matching, and change in metadata.

1.0 Introduction
The library community has been linking and mapping metadata before the linked data era. For instance, the testing project for Virtual International Authority File started in 1998, linking authority records created by international institutions. Catalogers can look up correlations between the Library of Congress Classification and Dewey Decimal Classification using Classification Web since 2004. Linking and mapping metadata can improve interoperability, lower maintenance cost, and increase the use of authority data both within and beyond the library community (OCLC 2019). Today, linked data shapes and supports the linking and management of metadata. It provides new approaches for authority control. For years, authority control has been text-based. Entities with the same name are differentiated through textual labeling and qualifiers. With linked data, we can manage identities by assigning unique identifiers, which is less language-dependent. While identity management does not replace text-based authority control, major metadata creators and aggregators like the Library of Congress and OCLC have been adding unique identifiers to entities to enhance their authority data. However, regardless of whether metadata carry unique identifiers, linking metadata is not without concerns. One challenge is that links between metadata may have different meanings, and the meanings may not be clear to users. A link can represent linking, which shows the linked metadata as related. This general sense of linking can cover many kinds of relationships (Green 2001). A link can also represent mapping, which often indicates a functional equivalence relationship between the linked metadata. When two terms are mapped, they are treated as equivalent functionally. That is, one term can stand in for another term. However, this does not guarantee semantic equivalence. An example is posting up a narrower term. In a standard, we may see USE cross reference that instructs users to use a broader term (e.g., dogs) for a narrower term (e.g., corgis). In this case, the two terms are treated as equivalent functionally, but not equivalent semantically. We can also see mapping between terms from standards with different levels of specificity. A link can also represent matching, which indicates semantic equivalence. Variant forms of the same Library of Congress Subject Heading (LCSH) is an example of matching. Since
these linking types may all be represented by the same expression: a link, without explicit specifications, the meanings of links may be ambiguous. Adding to the complexity of links, the distinctions I have outlined here, between linking, mapping, and matching are not always acknowledged and used consistently. People may use different categories for linking types, such as exact match, partial match, etc. Also, these linking types could refer to different semantic relationships (e.g., hierarchical, equivalence, associative), and there is no one-to-one relationship between a linking type and a semantic relationship. Recognizing the ambiguity of links, some projects link metadata with pre-defined relationships (e.g., DCMI metadata terms). This clarifies the meanings of links, but users may have to take extra steps to access the scope notes of the represented relationships.

Meaning changes over time is another concern for linked metadata. Assigning unique identifiers to linked metadata enables easier updates for the preferred form of an entity, such as name changes. Nonetheless, the updated form only represents the updated meaning of an entity. Users cannot trace the history of meaning changes or name changes of an entity. Without contextual information, users would not know which links of an entity were created before or after meaning changes, and whether the links were re-evaluated.

One other concern for linked metadata is unclear or inconsistent linking, mapping, and matching criteria. One example is a pilot project I observed in my ethnographic fieldwork. The project is an attempt of a group of librarians to explore mapping LCDGT (Library of Congress Demographic Group Terms) to LCSH. The group leader drafted mapping criteria. Members go through all LCDGT terms, and use the criteria to search for matches or closest matches in LCSH. In the mapping process, members surface different aspects of mapping, including concept, text string, and types of heading. When there is an exact concept match in LCSH for a LCDGT term, more complexities follow. For instance, the matched LCSH may or may not use the identical text string. Also, a LCDGT term may match with a variant form of a LCSH or a former heading. How could we distinguish and present the different types of exact concept match to users?

This project shows how sophisticated mapping criteria may be. When linking metadata, if we only show links between metadata without clear explanation of the criteria, we risk using a set of criteria that differ from users’ expectations.

Besides the concerns discussed above, previous studies such as the AAT-Taiwan project identified mapping and translation issues in developing the Chinese language Art & Architecture thesaurus (Chen, Zeng, and Chen 2016). Likewise, Zeng (2019) reviews research, standards, and projects, and discusses approaches to address interoperability issues in metadata mapping. These studies present categorizations of mapping issues and provide suggestions to address these issues. Building on previous research, this paper proposes a creative use of Ranganathan’s (1989) three planes of work. By extending its application from classification work to the linked data environment, we can use it as an analytical tool to discuss issues of linking, mapping, matching, and change in metadata.

---

1 The ethnographic fieldwork started in September 2015, and ended in November 2019. I shadowed a cataloger at an academic library to explore cultural influences in cataloging practices. Through participatory observations, informal interviews, and taking field notes, I captured rich cataloging scenarios of the cataloger’s applications of international and U.S. standards to catalog materials in various formats and languages, and her interactions with other librarians.
2.0 Cases

This section will apply Ranganathan’s (1989) three planes of work to three cases to analyze linking, mapping, and matching issues. The three planes are idea plane, verbal plane, and notational plane. Ideas are originated from the minds of their creators, and communicated through language. Language is the medium for the communication of ideas. The ambiguities of language (e.g., homonyms) are embedded in the representations of ideas. Recognizing the ambiguities of natural language in verbal plane, the notational plane represent disambiguated meanings or help arrangement. Notations may include language, symbols, and numbers (e.g., class numbers). The three planes provide a structure to break down the levels of abstraction of the concepts described in knowledge organization actions, such as classification and cataloging.

2.1 Eugenics in DDC, CCL, and NDC: discrepancies between the three planes

Tennis (2012) examines the subject ontogeny (i.e., the life of a subject over time) of eugenics in all editions of the Dewey Decimal Classification (DDC). The study demonstrates meaning changes of eugenics, and how those changes were reflected in the scheme over time. Using the three planes of work, we can view this as an example of changes in the idea plane (i.e., the definitions of eugenics) leading changes of the notational plane (i.e., the class numbers), while the verbal plane (i.e., the verbal expression “eugenics”) remains the same. Based on this study, I examine the subject ontogeny of eugenics in the New Classification Scheme for Chinese Libraries (CCL) in Taiwan and the Nippon Decimal Classification (NDC) in Japan, and bibliographic records with eugenics as a subject heading (Lee 2016). Through analyzing the titles and co-assigned subject headings in the bibliographic records, I capture different meaning changes of eugenics in CCL and NDC (Lee 2018). However, these changes at the idea plane were not reflected in the other two planes. The subject name eugenics remains the same, and the class numbers for eugenics in both schemes are relatively static. If we compare the ontogeny of eugenics in DDC, CCL, and NDC, we can identify two factors that may lead to issues in mapping metadata. First, the meaning of a concept may change and diverge in different languages and schemes. This makes mapping metadata in different languages more challenging. Further, when the idea plane is not in sync with the other two planes, there is a risk of misrepresentation and imprecise mapping. The three planes of work can help us identify subtle changes in the idea plane.

2.2 Vernacular title or non-Latin script title: change at the verbal plane

Through the aforementioned ethnographic fieldwork, I observed a case of change at the verbal plane. The cataloger Q [pseudonym], whom I observed in the field, retrieved a bibliographic record of a Japanese book using the library catalog. In the record, the title in the original script was labeled vernacular title, and the Romanized form of the title was labeled title. Q explained that the term vernacular is discriminative2. The East Asian Libraries and the cataloging community in the U.S. have stopped using this term for more than a decade. Q sent a proposal to change the name of this metadata attribute, and it was changed to non-Latin script title. This is an example of how the verbal plane

2 According to Merriam-Webster Online dictionary (2020), vernacular was first known and used as “using a language or dialect native to a region or country rather than a literary, cultured, or foreign language.” The origin verna means “slave born in the household.”
may shape people’s understanding and reaction to the idea plane of a concept. While both vernacular title and non-Latin script title may refer to the same concept, the nuance of an expression may carry different meanings and lead to different interpretations over time. Failing to account for this could cause issues in linking and managing metadata.

2.3 LCNAF and Wikidata: notational plane and structural interoperability

In the field, Q shared their observations of mapping between the Library of Congress Name Authority File (LCNAF) and Wikidata. When editing LCNAF records, catalogers can add links to other resources that describe the same entity. For instance, in the LCNAF for *Twain, Mark, 1835-1910* (Library of Congress 2020), catalogers can add a link to the Wikidata entry for *Mark Twain* (Q7245 2020). In this case, the two sources describe the same person, with the same verbal expression, using different notations.

The issue is, the LCNAF is identity based, while Wikidata is person based. For people who publish works using multiple identities, each identity has its own LCNAF record. The records of the same person are linked to one main record, which serves as a hub and links to different identities of the person. Wikidata collocates all identities of a person under one page. If we search for Samuel Clemens in Wikidata, we will be directed to the Mark Twain page. Hence, when linking a LCNAF record that represents one identity of a person to a Wikidata entry of a person with multiple identities, the link does not connect two records with the same scope. Mapping issues may occur even if the idea plane and verbal plane are identical. How could we present the structural distinctions so users do not assume an equivalence relationship between the records? How to clarify the meanings of links? If we access the LCNAF record for *Twain, Mark, 1835-1910* through the LC linked data service, we see the link to Mark Twain’s Wikidata page is under closely matching concepts from other schemes, which indicates a non-equivalence relationship. The Wikidata page for Mark Twain lists the Library of Congress authority ID for both *Twain, Mark, 1835-1910* and *Clemens, Samuel Langhorne, 1835-1910*. Users may infer the relationships between these identities from other metadata on the page, which specify that Mark Twain is also known as Samuel Langhorne Clemens. While both systems indicate the differences between the records, the meanings of the link are not explicitly clear. Could we improve this, maybe at the notational plane?

3.0 Conclusion

Ranganathan’s three planes of work in classification activity remains relevant in the linked data environment. By extending its application, we can use it as an analytical lens to examine issues of linking, mapping, matching, and change in metadata. Also, it accounts for meaning changes of the same concept over time, which may undermine semantic interoperability if not reflected in metadata linking. This complements studies that focus on linking issues between different metadata standards. Further, we can interpret and address linking issues by identifying discrepancies between the three planes. On one hand, following Ranganathan (1989), we expect the idea plane to lead the change of its expressions in the other two planes. Through the eugenics case, we see the risk of misrepresenting concepts when the three planes are not in sync. On the other hand, in the latter two cases, we recognize that change in the idea plane is not the only force of change for the verbal plane and notational plane. Change in these two planes, under the premise of remaining in sync with the idea plane, may help address linking issues.
References
Doi:10.1007/s00799-015-0163-1
https://doi.org/10.1002/asi.22686
Reconstruction of Crisis Knowledge Ontology by Integrating Temporal-Spatial Analysis

Abstract:
In this paper the restructuring of crisis knowledge by integrating temporal-spatial analysis was proposed and considered into ontology construction. It aims to explain the temporal-spatial features and composite effects of crisis knowledge, and then implement a fine-grained reconstruction model based on ontology.

1.0 Introduction
Ontology is one of the common methods of conceptual crisis knowledge organization because of its meaningful contribution onto the sustainable knowledge management and intelligent decision making (Mescherin et al. 2012; Matta et al. 2012). However, former researches for the ontology-based crisis knowledge organization were interested more in crisis types and attributes while ignoring the real-time dynamics of information during the emergency rescue (Moehrle 2012; Xu, Nyerges, and Nie 2013). Therefore, in this paper the dynamic organization of crisis knowledge will be analyzed and restructured relied on their temporal-spatial features by using ontology. The innovations of this study are as follows. First, temporal and spatial factors are segmented according to sequences and scenes, based on which the cross-effects of their dynamic changes are analyzed. Second, a fine-grained dynamic knowledge organization of feature units from the perspective of composite temporal-spatial relationships is proposed, which is supposed to express the dynamic semantic structure more accurately.

2.0 Reconstruction of fine-grained crisis ontology for integrating temporal-spatial features
This section will introduce how to reconstruct crisis knowledge by integrating temporal-spatial features. The research framework is shown as the figure below (seen in Figure 1):

![Figure 1: Research framework of reconstruction with integrating temporal-spatial features](image-url)
2.1 Analysis of temporal and spatial features of crisis knowledge

In this study we divide crisis knowledge into two categories: dynamic knowledge and static knowledge. Static knowledge represents the information that does not change during the whole development of the crisis, while the dynamic knowledge represents the information with characteristics of multi-phase and discontinuity.

Since the attributes of static knowledge are all fixed values, the corresponding temporal and spatial attributes could be simply labeled and added. While for dynamic crisis knowledge, we need to divide the time series into multiple time regions with short duration and high granularity (Pohl, Bouchachia, and Hellwagner 2018). Similarly, we divide the space into multiple scenes with small space and high density. The classification of disaster levels is applied to indicate the severity of disasters in a fixed area (Saoutal, Cahier, and Matta 2014). The correlation analysis explores the interaction between temporal and spatial characteristics, such as the division of time zones will affect the regional division of disaster levels. For example, in the same disaster area, the severity of disasters must be different in the crisis development time zone and the crisis recovery time zone. The integration research of temporal-spatial characteristics explores the influence of temporal and spatial characteristics on the organization of crisis knowledge. For example, we use the division of time zones and disaster levels to determine whether a region has been affected by a crisis and to what extent so as to determine whether this region is an area of crisis rescue or crisis supply.

2.2 Fine-grained ontology reconstruction based on temporal-spatial composite features

Combined with the analysis of temporal-spatial features, a fine-grained crisis knowledge ontology model could be implemented from the perspective of the temporal-spatial correlation. The temporal and spatial relationships represent respectively the temporal and spatial attributes of crisis knowledge; while the semantic relation indicates the crisis knowledge structure. The crisis knowledge ontology model integrating the temporal-spatial features analysis is shown in Figure 2.

Figure 2 Schematic representation of fine-grained ontology reconstruction
3.0 Conclusion

This study proposes a fine-grained knowledge reconstruction by analyzing the temporal-spatial composite features of crisis knowledge. It solves the problems of inflexible structure and lack of dynamic integration in crisis knowledge organization. Meanwhile it also suggests new methods and ideas for the ontology-based construction of crisis information resources and real-time crisis management, which could be further improved by knowledge graphs for example. Since it is work-in-progress research, the evaluation or instance verification of this ontology construction will be performed in our future studies.

References


Can an Ontologically-Oriented KO Do Without Concepts?

Abstract:
The ontological approach in the development of KOS is an attempt to overcome the limitations of the traditional epistemological approach. Questions raise about the representation and organization of ontologically-oriented KO units, such as BFO universals or ILC phenomena. The study aims to compare the ontological approaches of BFO and ILC using a hermeneutic approach. We found that the differences between the units of the two systems are primarily due to the formal level of abstraction of BFO and the different organizations, namely the grouping of phenomena into ILC classes that represent complex compounds of entities in the BFO approach. In both systems the use of concepts is considered instrumental, although in the ILC they constitute the intersubjective component of the phenomena whereas in BFO they serve to access the entities of reality but are not part of them.

1.0 The aim and scope of the study

In the literature associated with the Knowledge Organization area two approaches are found, the ontological one and the epistemological one, that are often associated with antagonistic positions, sometimes extreme (objectivism and subjectivism). Several authors, e.g.: Ridi (2016); Kleineberg (2017); Dahlberg (2017); Hjørland (2010), claim the lack of grounds for such positions, emphasizing that, in procedural terms, knowledge is both ontological and epistemological. As an area of study, ontology focuses on how reality is constituted and structured using concepts. However, even ontology uses concepts, the formation of which falls within the scope of epistemology, which studies how knowledge is acquired and grounded. Therefore, as Poli and Obrst (2010) state, the procedures of these areas complement and condition each other.

Seeking to overcome the inadequacy of the traditional epistemological approach in the development of knowledge organization systems (KOS) to interdisciplinary research, several authors argue\(^1\) that: “the basic units of the new KOS should be phenomena of the real world.” This turn towards an ontological approach raises questions about the representation and organization of these new units.

Based on an ontological realism, Arp et al. (2015) consider that ontologies shouldn't focus on concepts but on universals or types (bringing together common characteristics of instantiated particulars). Within this theoretical position Klein and Smith (2010) distinguish concept systems - as systems of meanings; from representation systems - which relate to real world entities. This is the approach of Basic Formal Ontology (BFO), an upper-ontology that bridges ontologies from different domains, recently adopted as an information exchange standard (ISO/IEC CD 21838-2).

\(^{1}\) During the 8th conference of the ISKO Spanish chapter (2017), as part of the proposals labeled as "the León manifesto." http://www.iskoi.org/ilc/leon.php (access on 8.1.2020).
Apparenty in a seemingly position of compromise, Gnoli (2018) includes concepts, as produced through the evolving interaction of percepts (individual experience of reality) and previously accumulated knowledge, in the notion of phenomena, that are taken as the basic units of the Integrative Levels Classification (ILC). The combinable facets of ILC include several “elements of perspective” seeking to integrate epistemological aspects into an ontological KOS as noted by Kleineberg (2020). This feature aims to overcome the limits of the disciplinary compartmentalization of KOSs such as the Dewey Decimal Classification or the Universal Decimal Classification which, in addition, present some misfit to the digital context as stated by Szostak et al. (2016); Dahlberg (2017); Simões et al. (2016); Gnoli (2008).

The aim of this paper is to compare the ontological approaches of the BFO and the ILC, by identifying the elementary units that constitute them and the guiding principles that underlie their organization. We discussed the preliminary results of an ongoing PhD research on this topic, hoping that our study will bring useful contributions, such as the systematization of a set of guidelines for ontological KOS modeling. We consider these contributions as being useful both in the scope of the bibliographic classification systems, where the ontological approach has less tradition, and in the scope of the computational ontologies where, as Machado et al. (2020) state, "a careful intervention can be the distinguishing mark of the KO and LIS areas," regarding a greater intervention of researchers from these areas in the development of this type of KOS.

2.0 The method

Methodologically, a comparison method will be employed using the hermeneutic approach complemented with content analysis methods following Bardin (2011) and Kuckartz (2014). The methodological complementary allows a more in-depth analysis of the literature that applies extended segments, the units of context, as a way to facilitate the understanding of the core units of analysis, the units of meaning.

3.0 ILC and BFO units

ILC (2nd edition) units — the phenomena apprehended from reality, are related in six levels (strata) of formal dependence (information; matter; life; mind; societies; and works). These are divided into layers that are materially dependent on each other (e.g., the life stratum has, in order of dependence, four layers: genes; bacteria; eukaryote organisms; populations). The layers (25 in total) form the main classes of the ILC which, in turn, contain subclasses in variable numbers (e.g., fungi, plants and animals are some subclasses of organisms). Subclasses are ordered (whenever possible) by the principle of evolutionary appearance (emergence), each forming its hierarchical chain of types and subtypes (e.g., animals> chordates> mammals> whales).

It is in this latter form of organization, in a taxonomy of types and subtypes that the BFO units — the entities of the real world, are structured. Being a formal ontology (in the Husserlian sense of being neutral to any domain) its units are very general types of entities. The first division is between entities that continue or persist through time (continuant) and entities that unfolds themselves in time (occurrent). In BFO 2.0 continuant has three subtypes: independent continuant (with two subtypes of entity: material and immaterial); specifically dependent continuant (also with two subtypes: quality and realizable entity) and generically dependent continuant. As subtypes of
**occurrent** BFO presents four: *process*; *process boundary*; *temporal region* and *spatiotemporal region*.

Given the not entirely taxonomic structure of ILC, mapping between ILC layers of phenomena and BFO entity types should be understood as a mere academic exercise for the comparative study of its ontological structures. An attempt to distribute the ILC layers across the three main types of BFO entity (*independent continuant*, *dependent continuant* and *occurrent*) is shown in Table 1. In this comparison we do not include the ILC layer *forms*, belonging to the *information* strata, since this class deals with mathematical entities and, according to Smith et al. (2015), BFO 2.0 does not deal with those entities, although they are partially addressed in lower level ontologies built on BFO.²

Table 1: Exploratory distribution of ILC classes by types of entities at the top-level of the BFO strata

<table>
<thead>
<tr>
<th>ILC strata</th>
<th>ILC layers</th>
</tr>
</thead>
<tbody>
<tr>
<td>information</td>
<td></td>
</tr>
<tr>
<td>matter</td>
<td>energy; atoms; molecules; celestial bodies; rocks; land</td>
</tr>
<tr>
<td>life</td>
<td>genes; bacteria; organisms; populations</td>
</tr>
<tr>
<td>mind</td>
<td>instincts; consciousness</td>
</tr>
<tr>
<td>societies</td>
<td>language; polities; technologies; rituals; communities; enterprises</td>
</tr>
<tr>
<td>works</td>
<td>artifacts; knowledge; artworks</td>
</tr>
<tr>
<td></td>
<td><strong>BFO independent continuant</strong></td>
</tr>
<tr>
<td></td>
<td><strong>BFO dependent continuant</strong></td>
</tr>
<tr>
<td></td>
<td><strong>BFO occurrent</strong></td>
</tr>
</tbody>
</table>

The option for the highest level’s types of BFO units is due, on the one hand, to the inclusion, in some ILC classes, of phenomena that may be considered entities of different types at more BFO specific levels and, on the other hand, due to the ontological complexity of some phenomena involved. Still, for some classes the relationship of the respective phenomena to more specific types of BFO entities appears to be less intricate. This is the case of classes included in *matter* and *life* strata as well as of the *artifacts* class. Except for *continuum bodies*, which contain phenomena that can be considered within the *quality* type, the rest will be included in the *material entity* type. As for the class *energy* we follow the “future directions” of BFO as pointed by Smith et al. (2015, 6): “[p]ortion of energy potentially to be treated as child of material entity.”

The remaining ILC classes in the *dependent continuant* column of Table 1 contain phenomena that do not fit into the *quality* entity type, such as *continuum bodies*, but depend specifically or generically on the existence of some material entity to exist. Some could be considered what Smith (2016) calls a "total complex of dispositions" of which, according to the author, the English language, the law, the price system, the Christian

---

² The Information Artifact Ontology and the Ontology for Biomedical Investigations provide the resources to deal with numerical measurement results and with certain other mathematical entities.
faith are examples. In any case, when it comes to "social objects," the ontological analysis becomes more complex.

Finally, the phenomena of the two classes of the information stratum would be of the entity type spatiotemporal region. The spacetime layer will be the only one at the formal level of the BFO entities.

4.0 Discussion

When comparing the units of the two systems the most obvious difference can be attributed to the purely formal level of abstraction of the BFO. This characteristic is decisive for the gap between the units of the two systems. Another relevant difference is the grouping of phenomena in some ILC classes which, in the BFO approach, represent complex compounds of entities. In technologies class, for example, we can find the technological systems themselves, that would be a type of dependent continuant, and occurrent entities such as fishing and mining. Nevertheless, besides the spacetime class alignment, another similarity can be pointed out. One of the reasons for mathematical entities not integrating BFO is that they stand outside of space and time. In ILC, the class forms, that encompass those entities, is the only one that is not included in the nature “superclass” that encompasses, besides the other classes of the information strata, all classes belonging to the matter, life and mind levels.

Another similarity between the two systems will be in terms of the instrumental use of the concepts, although with different theoretical and, allegedly, procedural positions. In ILC, concepts are understood as units of meaning that, validated by science, constitute the intersubjective component of phenomena. These will be located in the ontological dimension that can be publicly represented. In BFO the concepts are used in the Aristotelian tradition of ontological analysis. In this perspective the ontology represents the real-world entities, and their relations, and not the concepts used to access them. These are understood as being meanings agreed upon by a group of responsible people and, in the words of Klein and Smith (2010, 437), "'partitive' and 'associative relations' (such as part_of or causes) are not appropriate for concept systems - since they are not relations which hold between meanings." For those authors, only the narrower_than relationship and its inverse should be used to link concepts. One may wonder which model is the most consistent in the ontological approach to knowledge organization.

5.0 Final remarks

The initial state of the study leaves several questions open that we intend to address in the future. Among these, is the deepening of the relationship between conceptual and representation systems, the nature and constitution of concepts and the distinction between epistemological and ontological knowledge. We consider them to be complex issues, but of great relevance for an area of study that intends to organize knowledge.

References


Enhancing Knowledge Organization Through Implicit Collaboration in Crowdsourcing Process

Abstract:
This paper presents our approach in removing noisy labels from crowdsourced data and enhancing understanding and communication through crowdsourcing process aimed at creating metadata for describing digitized artworks of a University Library Museum. A responsive Web application was created for the crowdsourcing activity and made open to the University community for interested individuals to participate in annotating the images. The collected annotation in form of tags were preprocessed and filtered to generate a subset of tags by removing duplicates and also eliminating some noisy labels using majority voting. The resulting subset was used as labels for the images for a second round of crowdsourcing process where users chose from the filtered labels. Comparing the output of the second round with the label (tags) from an expert shows a high level of similarity between the selected tags and the expert generated tags.

1.0 Introduction
People read different meanings to images and view them from different perspectives. The creator of an image always have an intended information to communicate and users of the image interpret the image based on their existing knowledge and also the context in which they find themselves. Images like artworks are more difficult for users to interpret and most often need explanation by an expert. In museums, most of the collections have labels and a little description of the work in order to assist visitors (users) to relate with the collections. Despite this, users still find it difficult to understand some of the collections and, most often, need a guide to provide them with explanation and to answer their questions. When these collections are digitized and rendered online, it becomes more necessary to help users in interpreting the artifacts. Several approaches have been used in the literature. Sharma and Siddiqui (2016) developed an ontology based approach for retrieval of digitalized museum artifacts. They claimed to support semantic retrieval by automatically extracting ontological concepts in form of visual and textual features from images and their textual descriptions. Some other approaches involved the use of experts in creating ontology. However as earlier stated and also supported by the work of Zhitomirsky-Geffet et al. (2016), “experts are able to build knowledge organization schemes and ontologies of high professional quality, but experts are hard to find and expensive to employ”. The other problem related to this is the non-scalability of the ontologies as they are domain and context specific and for new collections, new ontologies would have to be created. Also, the advancement in Internet and Web technologies, and the need to disseminate artifacts to a wide audience has led to a continuous rise in the digitalization of existing artifacts and creation of new digital artifacts. However, human experts cannot adequately meet up with the challenge of providing prompt knowledge organization service to cater for this rise.

In information retrieval research, the limitation of taxonomy, expert indexing and ontologies have been discussed severally. Approaches like social bookmarking, social indexing, and collaborative tagging have emerged and have been extensively explored.
to bridge the gap between controlled vocabularies used by experts in classifying documents and natural language expression of information needs by users (Golder and Huberman 2006). Users’ participation through social indexing and collaborative content creation have enhanced knowledge organization and information retrieval despite the inherent problem of “noise” associated with users’ content in knowledge organization activities.

Using non-experts in knowledge organization can be considered as a crowdsourcing process. Crowdsourcing has been defined as “the process of bringing in many people to achieve great feats from tasks that used to be handled by only a specialized few” (Howe 2006). Crowdsourcing has been applied to several areas like information search, image labelling (Jackson et al. 2018), data classification, document translation, sentiment analysis, organizational learning (Lenart-Gansiniec and Sulkowski 2018) and many more. The shortage of experts (“specialized few”) in handling astronomically increasing documents generated on daily basis can be considered as one of the reasons for increasing resort to crowdsourcing.

Crowdsourcing of data come in two different forms. The first approach involves the requester providing a list of labels (or answers) from which the participants can choose from. The second form is to allow participants to type in labels (or answers) based on their perception of the object on which data is sought (Adeogun and Odumuyiwa 2019). Crowdsourcing is a social process which can also be referred to as implicit collaboration. This process comes with the advantage of enhancing speed in solving problems by depending on the wisdom of the crowd. It can also culminate in collective intelligence. It is however important to point out a fundamental problem of “noisy labels” prevalent in crowdsourced data (Adeogun and Odumuyiwa 2019). Since the participants are not necessarily experts in the domain, provided data may be riddled with impreciseness and unnecessary tautology especially when participants type in their responses freely without any constraint of choosing from a list of labels.

2.0 Methodology

The objective of this work is to explore how to enhance the representation of the University of Lagos collection of artworks. The goal is to explore the possibility of engaging users of the artwork in generating tags to describe what the artwork depicts. Of course, describing the artwork can be done by an expert or some experts in cultural artifacts. However, experts are scarce, their services are expensive and they are limited in number. They are more suited for collections of smaller number. The observed increase in the creation of digital artifacts and contents exposes the limitation of having enough experts to manually provide metadata on digital artefacts. The University of Lagos Library has a museum that contains over 123 artworks some of which have been displayed in international expositions. Two of the artworks are shown in Figures 1 and 2. The University Library has digitalized the artifacts and would like to create an online information retrieval system for the content.
We created a responsive web application (see Figure 3) to crowdsource for tags and comments on the images of the artifacts. In the first phase of this research, a total of 28 artifacts were posted on the application. Participants (users) were drawn from the University community by broadcasting the link to the Web application and allowing interested individuals to create an account on the application. When the account is created, the users log in with their account credentials and start viewing the images and providing annotation in form of tags and comments. Our objective is to capture users understanding of the artifacts without any aid hence we did not provide any label or description. Users could add as many tags as they want to any of the images. Running the application for just one day, over 25 participants subscribed and generated over 430 tags. Each image received an average of 15 tags.
We preprocessed the data to correct some typographical errors and also did a frequency count of each tag expressed on an image. With the processing, the tags were filtered and we selected about 10 most occurring tags for each image and we modified the application using the selected tags as labels for the image. Another call for crowdsourcing was made and about 20 users responded. In this second call, the application was modified such that users were asked to select the 3 options that best describe each of the images (see Figure 4). About 20 users responded to the second call with about 671 tags selected in total. We applied majority voting on the new tags to remove noisy labels and select the best three for each image. We also engage an expert librarian focused on artworks to provide metadata for the images and we compared his output with the output of the crowdsourcing process as shown in Table 1.

3.0 Result and Discussion

We compared the result of the crowdsourced process with that of the expert and observed high similarity among the two. The essence of this work is to test our approach in removing noisy labels from crowdsourced data by experimenting with few artworks. Of course the end goal is to apply this approach to a large corpus of artworks with or without input from experts. The initial results we obtained in this work convince us that with large corpus of artworks, crowdsourcing will make it faster for us to obtain relevant metadata to describe our artifacts and would enhance the information search process using the information retrieval system to be created.
Table 1: Comparison of the crowdsourced tags with the expert tags for four images

<table>
<thead>
<tr>
<th>Crowdsourced Tags</th>
<th>Expert Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graduation</td>
<td>Matriculation</td>
</tr>
<tr>
<td>Arts</td>
<td>Cattle rearer</td>
</tr>
<tr>
<td>Mahatma Gandhi</td>
<td>Wise man</td>
</tr>
<tr>
<td>Hunger</td>
<td>Poverty</td>
</tr>
</tbody>
</table>

4.0 Conclusion

This is an ongoing research work aimed at creating an information retrieval system on artworks. This paper reports the first phase of the research focused on crowdsourcing tags from the University community to serve as metadata for describing artifacts. We used the majority voting approach to remove noisy labels from the crowdsourced data. It was observed that the resulting data were very similar to the data provided by an expert on the artworks. In addition, we observed that more semantically relevant tags were generated during the crowdsourcing process. This can enhance the work of the expert in providing metadata on the artifacts. We plan to improve the result by using some other algorithms for removing the noisy labels without having to reduce user selections in the second round of crowdsourcing process.

References


Jackson, Corey, Kevin Crowston, Carsten Østerlund, and Mahboobeh Harandi. 2018. “Folksonomies to Support Coordination and Coordination of Folksonomies.” Computer Supported Cooperative Work (CSCW) 27, nos. 3-6: 647-678.


Lucia Maria Velloso de Oliveira – Universidade Federal Fluminense, Brazil  
Bianca Therezinha C. Panisset – Universidade Federal Fluminense and Fundação Casa de Rui Barbosa, Brazil  
José Antonio da Silva – Universidade Federal Fluminense and Fundação Casa de Rui Barbosa, Brazil

Mediation in Archives  
Organization, Classification and Transparency

Abstract:  
It discusses the role of mediation in archives and the use of technology in supporting the process. As a central issue we investigate how the organization of the archives is related to a set of functions that aim, in the final analysis, to grant the user access and, in the case of collections of historical value, to enable future generations to access the past. The general objective is to present research results that relate to three pillars of the process developed within the scope of the information units, with an emphasis on the archival context: organization, classification and transparency. As a method, we use the analysis of a site of historical value and a governmental adherence process for the implementation of digital administrative processes. Our result shows that these initiatives were only possible due to the role of mediation in archives as a pillar for the organization of knowledge.

1.0 Introduction  
This paper aims to present research results that relate to the three pillars of the process developed within the scope of information units, with an emphasis on archival context: organization, classification and transparency. Our premise is that technology is a tool to democratize access to documents and promote identity, memory and transparency. In this perspective, the work is developed in the area of records management and in the area of dissemination of collections relevant to the production of knowledge about a historical period. In both situations, the use of technology is essential to democratize access to records and to promote identity, memory and transparency. Firstly, the analysis focuses on the implementation and maintenance of a website designed to promote collections about slavery and abolition in Brazil, one of the countries that had the largest flow of slaves in the world. The site aims to attract new users of archives relating to these historical events and thus render the archives something more real and part of people's imagination and life. Secondly, the analysis is empirically based on the implementation of a project aimed at the registration and use of digital documents in public management. The project involves Brazilian federal agencies, since they were included in the broad plan for the implementation of electronic processes in the structure of public power that had as an objective the transparency of this management.

2.0 Organization, Classification and Transparency as approaches in highlighting the archival context  
There are two archival functions that strengthen access, and, of course, transparency and the right to memory. The organization of archives is related to a set of functions that aim, in the final analysis, to grant the user access to the acts of an organization and, in the case of collections of historical values, to enable future generations to access the past. In the field of archives, classification occupies a privileged place in both theory
and practice. It is through classification that the organic relationship of archival documents and their links is established and revealed. Organicity, according to Luciana Duranti, is at the heart of archival science: “At the core of archival science is the concept of archival bond, that is, the network of relationships that each record has with the records belonging in the same aggregations” (Duranti 1997, 216). Organicity is understood as the connections existing between documents, established since their production and determined by the function exercised by each document in the context of representing an organic activity of an individual or an institution. We also use Yeo's understanding of archival representation. He states that: “...the concept of record representation means that records can be characterized as a persistent representation of activities that are created by participants or observers of those activities or by their authorized proxies” (Yeo 2007). For the author, records are ultimately “persistent representations of activities, created by participants or observers or their authorized proxies” (Yeo 2007).

The research carried out assumes that the archival document is a unit of the Knowledge Organization (KO), as mentioned by Hjørland (2008, 96), but also a persistent representation, which reveals not only the activity and the reason why it was produced, but the context of its production: actors, institutions and archival context. This understanding of the concept of archival context is based on the formulation of Thomasen (2001), which recognizes determining factors beyond the context of production, such as the context of use and the socio-political, cultural and economic context. Thus, the classification of the archival document, in the traditional or digital environment, involves the physical and intellectual organization of the records, so that their relationship with their genesis is represented, functioning as a mechanism for the recovery of the organization of information and knowledge (KO), so that it can be used as a framework for defining record deletion. We are then starting from the understanding that in the process of organization and classification in the field of archives, we deal with knowledge about the producer: its functions, business processes, relations and the historical context in which it lived. In addition, this knowledge will guide the relationship with the users of information systems and will be used as a source for the promotion of government transparency, so that accordingly at any time the acts will be understandable to society.

According to Pekka Henttonen (2015) in KO contextual classifications, concepts and relationships are linked to functions, activities and actors in the environment of the production and use of records. Archival records, as we know, despite being contextualized in their genesis according to their functions, can be classified according to their content, when the question is access, as it is the analysis of the content that will give the crucial elements for defining the degrees of confidentiality and consequently justifying access restrictions that impact on government transparency. As for transparency Meijer (2015) understands that “Many analyzes of government transparency focus on new and even future developments [...] These analyzes are useful to understand current issues, but generally fail to put them in perspective (historical)”. It is observed, therefore, that in the field of Archivology a better combination of actions aimed at promoting knowledge is necessary with regard to the context and the relationship between process and technology. To clarify: a process corresponds to a set of joint activities that aim to achieve an objective. Technology, in turn, consists of means and techniques to facilitate or streamline this process. In this sense, for the technological activity to be oriented in favor of the objectives of a process, instruments recognized in areas of knowledge are
used whose premises collaborate with the purpose of the technical-procedural relationship. An example of this possibility is in Document Management, which can be considered a macro process with several steps necessary to fulfill the archival functions, among them production control, classification, organization and destination of documents. We analyzed two technological environments applied to archives that, in our understanding, allow and enhance the process of transparency, recovery and informational access: the Slavery, Abolition and Post-abolition website and the implementation of the national electronic process. In both cases the information or digital copies of documents were classified and organized in their origin, and this was repeated in the digital environment.

3.0 The diffusion of historical collections and the management of records

The Slavery, Abolition and Post-abolition memory website is a virtual environment that brings together digital objects and a space for interaction with its users about the themes surrounding slavery, the process of abolition of the enslaved and the period after abolition. This environment presents society with a variety of materials on its theme: digitized documents, images, controlled vocabulary, links to related websites, documentary exhibits and four types of games aimed at young audiences.

The Brazilian public institution responsible for maintaining the site is the Fundação Casa de Rui Barbosa (FCRB), which offers a space reserved for intellectual work, consultation of books and documents, and the preservation of national memory. The website was inaugurated in November 2015, with the objective of providing the dissemination of historical collections, attracting new users to the archive, enhancing access to custodial documents relating to this the subject, bringing society closer to this archival collection through applications, games and document samples and to promote the process of identification and construction of memory triggers (Oliveira, 2014). The project involved: “Translation into Portuguese of The Trans-Atlantic Slave Trade Database from Voyage portal [...] ; exhibition catalogs; educational module (texts and games); elaboration of a controlled vocabulary about Afro-descendant memory; and providing a link library” (Oliveira 2014). The table below presents the data on the systematization of user service and the accesses transmitted through the use of the website that occurred between the years 2016 and 2019, where we obtained definitive numbers. It should be noted that the dissemination of the project took place in academic-scientific events, online on the Foundation's website and in person when the Archive Service received specialized technical visits from professionals and students.

<table>
<thead>
<tr>
<th>Table 1. Systemization of site use from 2016 to 2019</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of users</td>
</tr>
<tr>
<td>Number of initiated sessions</td>
</tr>
<tr>
<td>Digital bookshelf</td>
</tr>
<tr>
<td>Exhibitions</td>
</tr>
<tr>
<td>Controlled vocabulary</td>
</tr>
<tr>
<td>Related sites</td>
</tr>
<tr>
<td>Educational module</td>
</tr>
</tbody>
</table>

1 http://www.memoriaescravidao.rb.gov.br/
2 Source: Google Analytics. Nnumber of initiated sessions refers to the interaction over a period of time.
In this project, the necessary alignment between process and technology is observed because, for the site to exist with the mentioned content, it was necessary to plan beforehand involving, in the first instance, an adequate classification and organization of the documents whose praxis had occurred within the scope of the Historical and Institutional Archive of the Casa de Rui Barbosa Foundation. The area of games that aims to create young people's interest in their history and in the documents that translate their history is one of the most visited and has been successfully tested with young people from public schools, and presented at the BIG Festival, which is a festival of independent games from Latin America. In the same way, we can verify another action of conjugation between technology and process in the scope of the management of electronic processes. Brazil has adopted a government project entitled National Electronic Process (PEN) “a joint initiative of agencies and entities from different spheres of public administration to build a public infrastructure for the electronic administrative process” 3. For the scope of Brazilian state institutions, it was decided to adopt the “Electronic Information System” - SEI, which allows the creation and processing of administrative processes through digital means. The implantation of SEI at Fundação Casa de Rui Barbosa took place under the leadership of an archivist and followed the parameters of the institutional Document Management Program, something that was not repeated in other public institutions, whose process of adhesion and implementation of the program was led by the Information Technology area (Panisset and Jaccoud 2019). The system allows, in the assessment of the administrative process, the application of archival classification, although it does not have any archival control module for document management, which is performed outside the software.

As advantages, we point out two aspects mentioned by the Brazilian government: “increasede publicity of the processes, making it easier for them to be followed by civil servants and administrators, and their internal control and by society; and expansion of knowledge management and the possibility of improving processes, due to the creation of a single platform that allows the analysis of process flows, their comparison between different bodies and improvement based on successful experiences” 4. This system could not be adopted without the necessary prior measures, including a review of the procedures for organizing and documenting classification, adapting, in the computerized context, to the successful practices within the scope of the analog process already implemented by the archive area. These measures, in turn, collaborate on the access to information and, consequently, for the transparency of public administration acts, as indicated by the above mentioned advantages.

In this context, there is an expansion of access and, concomitantly, the importance of the correct management of archival documents is signaled as a necessary process for the new technological imperative, collaborating in the promotion of knowledge in the historical perspective (identity and memory) and in the political-administrative perspective (access to information and transparency of public information). In the area of archives, there is this same contribution, because the management of documents culminates in the

---

4 processoeletronico.gov.br/index.php/assuntos/produtos/sistema-eletronico-de-informacoes
decision on whether or not documents remain for a fixed time or for life. These documents whose timeless permanence is generally necessary represent milestones for the identity of a people and the memory of a nation.

4.0 Final considerations
This discussion considered the use of technology as a mediating element in the information access and transparency process. We presented two projects developed by a Brazilian public institution where technology has resulted from the premise that technology is a tool to democratize access to documents and promote identity, memory and transparency. It is the organizational activity that encompasses all the actions necessary to guarantee access, which will either promote the forgetfulness or the visibility of the records. The application of a technological solution for the production, retrieval and dissemination of information dispenses with the mediating action in the collection, aimed at classification, organization and transparency. The analyzed applications (SEI and thematic site) were preceded by the process of elaborating the archival arrangement and description in the case of the Slavery, Abolition and Post abolition memory site and also the previous implementation of a document management program, in the case of the application of software for the prediction and processing of administrative processes. These actions, which are not limited to this sample scope, aim to meet the different uses of the document: evidence, identity, memory, access and transparency. Therefore, we observed that the use of the document and the interest in the information contained therein is broad and can be applied in different approaches, but that mediation is necessary for the better use of the organization of knowledge.

References
Abstract:
In academic contexts, students and researchers are expected to follow a set of strict conventions that are strange or unknown to them, especially when their dominant language is not English. This paper aims to show how tools such as the Corpógrafo (Maia and Sarmento 2005), a web-based environment for corpora research (DIY - Do it Yourself corpora), can be used to provide writing assistance, in this particular case on the domain of Knowledge Organization (KO). To conduct this study, we used a corpus of articles published in Knowledge Organization, the official bi-monthly journal of ISKO founded in 1973. The corpus is composed of articles published between 2016 and 2018 in the sections ‘Articles’ and ‘Reviews of Concepts’ in Knowledge Organization. The articles were processed in Corpógrafo, in order to enable results to be extracted. Given the results obtained, the relevance of using Corpógrafo as a writing assistance tool for the KO scientific field is discussed, in particular in the areas of: (a) terminology extraction; (b) retrieval of collocations and identification of specialized language phraseology.

1.0 Introduction
In academic contexts, students and researchers are expected to follow a set of (more or less) strict writing conventions that are often alien to them, especially when English is not their dominant language. This demands a command of both the general language (LGP), i.e. is the language used in routine daily interactions, and of the Language for Special Purposes (LSP), used to exchange and disseminate specialized knowledge, and hence facilitate specialized communication. Drafting texts in LSPs thus demands a command of the general language (as any LSP builds upon the general language grammar and lexis), but primarily of the specific conventions of that particular LSP. Appropriate information searching is thus crucial to assist users, as it allows them to produce texts using correct conventions, as well as the right linguistic form and contents. The main advantage of writing assistants, to which a significant volume of research has been devoted (e.g. Bourelle, Bourelle, and Rankins-Robertson 2015; Tarp, Fisker, and Sepstrup 2017), is that they offer novice and technical writers the support to produce texts on a particular topic. Therefore, LSP texts show some particular features: (a) they are terminology-rich (i.e. they contain a high density of specialized terminology, including formulae, acronyms and abbreviations, among others); (b) they contain particular word combinations (collocations); and (c) they use grammar differently from LGP (e.g. simple and short sentences, impersonal language, few adjectives, and word repetition to the detriment of co-referencing in an attempt to preserve accuracy, concision and precision, while avoiding ambiguities). Thus, tools that are commonly used in general language writing assistants, such as word processor spell and grammar checkers, can be useful to get the general language spelling and grammar right, but will hardly help them produce a text in a particular LSP.

There has been a growing interest in DIY (‘do it yourself’) corpora for LSP. These are composed of small to medium scale collections of texts, whose choice relies entirely on the user's specific needs within a given discipline (Maia 1997). The features offered by corpus linguistics, such as word lists and concordances, are particularly useful, as together they can show the most frequent words and word collocations, provide
information about the meaning of a word, detail the use of a word when combined with (an)other word(s), and provide a context for a certain word. Arguably, thus, DIY corpora can be used as a writing assistant, hence serving as a tool to improve LSP writing skills. This paper addresses writing assistance as a methodology, more than as a tool.

2.0 Objectives and methodology

Education for Knowledge Organization (KO) is a core, especially in LIS courses (Alajmi and Rehman 2016), including in countries where students are not native speakers of English. With the primary aim of providing students with a tool that allows them to explore specific vocabulary/terminology of KO, tests were run in the 2019 version of Corpógrafo (V5)\(^1\), a web-based suite of tools for collection and analysis of DIY corpora. This suite, which is freely available, allows text collection, corpus search (search for regular expression concordances, extract collocations, and produce frequency-based statistics - such as \(n\)-gram counts), information extraction (collect terminology, establish semantic relations, produce conceptual maps); knowledge-resource building (collect specific-domain glossaries, thesauri, terminological databases and ontologies), and categorized word-lists (Maia and Sarmento 2005). The tool has already been used for conceptual studies in the field of information science (Oliveira and Rodrigues 2011).

The tests run allowed the study of the linguistic specificities in KO, with a view to further its linguistic enquiry and the analysis of the evolution of this scientific area. In this context, the following features of Corpógrafo were used: (a) terminology extraction; (b) retrieval of collocations and identification of specialized language phraseology. The corpus used in this study builds on articles published in Knowledge Organization, the official bi-monthly journal of ISKO founded in 1973. It consists of 16 articles published between 2016 and 2018 in the sections ‘Articles’ and ‘Reviews of Concepts’ in KO. The terminology extraction was performed using the Terminology Database (TDB) feature, whereas the collocations and phraseology were obtained using the concordances search functions (KWIC, sentence search and window search). In order to establish the frequency counts of the words, we undertook a study of \(n\)-grams and thus have an overview of the text contents. This analysis allows the visualization of some prominent features of the corpus, e.g. systematic use of certain words or word strings that enable the identification of terms, collocations or writing styles. Finally, a list of the most frequent terms was created, which was used as the basis to select terms for retrieval of collocations and identification of specialized phraseology.

3.0 Results and discussion

The results presented in this section are based on searches performed over the KNOWA - Knowledge Organization Writing Assistant corpus (our own DIY corpus of highly specialized KO publications) using the Corpógrafo. Overall, these results show that the methodology proposed has a significant potential as writing assistant. A list of \(n\)-grams was produced in order to establish which words/word combinations are most frequent. Linguistic elements that were not relevant for the KO LSP were discarded. We manually selected 5 terms from the word list produced by Corpógrafo that are core to KO and often used to build more complex terms, based on two criteria: (1) they are

\(^1\) Available at https://www.linguateca.pt/corpografo/
highly technical terms; (2) they are some of the most frequent terms in KNOWA. These are: ‘classif’ and ‘index’ (using the * wildcard); ‘concept’; ‘domain’; and ‘subject’.

3.1 Terminology extraction

Terminology, like collocations and specialized phraseology, can also be extracted from highly specialized language corpora. Corpógrafo includes a terminology extraction function that is particularly helpful, as it returns a list of candidate terms, which the user can then choose to include in the terminological database, or else simply discard (in case they are not terms). As the terminology search function in Corpógrafo is based on user-selectable search strings (from 1- to 5-grams), it is likely that the candidate terms include words that are not part of the term. While the strings containing only general language should be discarded, those strings containing terms should be included in the Terminology Database (TDB) and edited to remove unnecessary words. Nevertheless, as we consider that (a) not all corpora management software enables terminology extraction, (b) not all those that do operate in the same way, and (c) Corpógrafo is localised in Portuguese, in this research we use simple n-gram searches, followed by human selection of terms, to illustrate how corpora management tools can provide resourceful writing assistance. The following terms, sorted alphabetically, are illustrative examples: “Aboutness is a concept used in LIS”; “work domain analysis in CWA (like DDD) differs from domain analysis”; “Formal Concept Analysis (FCA)”; “Indexing”; “Library of Congress Subject Headings (LCSH)”; “MeSH (Medical Subject Headings)”; “The alternative principle is request-oriented indexing”; “mainstream automatic indexing is not purely document oriented”; “subject access points”.

This list reveals that there are one-word terms that are core to KO (e.g. ‘indexing’), as well as multi-word terms that show different possibilities within a broader term (e.g. ‘request-oriented indexing’ and ‘mainstream automatic indexing’). When writing about information science it is crucial to know that there are ‘subject access points’, as well as different subject headings, including ‘Library of Congress Subject Headings (LCSH)’ and ‘MeSH (Medical Subject Headings)’. Given the requests for precision and concision underlying terminology (Cabré 1999), it is essential to consider that terms are often expressed in shortened forms (including acronyms and abbreviations), and defined in the text. The list of terms above shows several examples of the former: ‘LIS’, ‘FCA’, ‘LCSH’ and ‘MeSH’. Use of these shortened versions, rather than the full versions, will show that the writer has a command of the terminology in the area.

3.2 Retrieval of collocations and identification of specialized phraseology

Retrieval of collocations from corpora has a significant potential for providing writing assistance to inexperienced writers. Information on the collocations of a certain word will steer the writer, by providing them with lexical combination possibilities. By resorting to this method, users can do searches similar to the ones provided by Benson et al.’s (1986) resource, with the significant advantage that they have access to an updated, more easily and faster search capability - and, most importantly, to a resource that applies to their LSP and not to LGP. Word collocations also help writers select the right option from a taxonomy of possible options. Additionally, those collocations allow the writer to learn from specialized writers how to use field-specific phraseology. This is crucial since, similarly to what happens with general language — where a non-native
speaker is identified as such by their use of uncommon collocations and wrong prepositions — the use of uncommon collocations in LSPs may indicate lack of proficiency in the field. The following table illustrates examples of collocations and phraseology for the 5 terms mentioned above.

Table 1. Examples of collocations and phraseology for the 5 terms mentioned above

| Classif* | Indexes may be classified according to / Objects may be classified based on / systems of classification and categorization / the classification of knowledge |
| Classification [systems] | Dewey Decimal Classification (DDC) / Universal Decimal Classification (UDC) / Library of Congress Classification (LCC) / Bibliographic classification / reader-interest classification / Aristotelian classification |
| Concept | instances of the concept / representation of a concept / practice of concept systems construction / knowledge-based approaches to concept formation |
| Domain | creation of the domain / domain knowledge organization / develop the domain |
| Index* | processes of cataloging, subject analysis, indexing and classification / alphabetical index / the development of citation indexes / indexing principles / documents are being indexed / the activity of assigning index terms |
| Subject | lists of subject headings / subject knowledge / subject fields / subject retrieval / subject classification schemes / LIS specialists assign subject labels to documents / What is the criterion that a given subject should be attributed to a given document? |

These examples illustrate the potential of corpora management software as writing assistants. A writer can learn, after performing a corpus search of the stem ‘classif’ + the wildcard * (in order to obtain all derivatives of ‘classif’), that indexes are classified according to a set of criteria, whereas objects are classified based on other criteria. We also learn that systems of classification and categorization often work closely together, and that knowledge is subject to classification, not grouping or systematization. A corpus search for the same stem also unveils a taxonomy of classification systems: Dewey Decimal Classification (DDC); Universal Decimal Classification (UDC); Library of Congress Classification (LCC); Bibliographic classification; reader-interest classification; Aristotelian classification. Likewise, when searching for the word ‘concept’, one learns that: there are instances, not categories or examples, of a concept; a concept has a representation, not a formulation; concept systems are constructed, not built; concepts can be formed - not created - using knowledge-based approaches. General language speakers and dictionary users will know that a domain refers to an area or territory owned or controlled by a particular power, but will be unaware that domains, in KO, can be created or developed, but not produced.

One crucial term in KO is ‘index’ and its derivatives. The concept is so complex that less experienced writers on the topic can struggle to match the right terms and concepts. Again, a corpus search will be useful to reveal that there are several processes that are approximate: cataloging, subject analysis, indexing and classification. By being aware that these terms are used to name close concepts, writers can look for the fine detail that set each concept apart. It will also show them that indexes can be alphabetical; there are citation indexes, and these are developed; indexing follows a set of principles; documents are indexed; and index terms are assigned. Finally, when writing about ‘subject’,
users might like to know that subjects may have headings and lists can be created from them; there is subject knowledge, but not subject comprehension; subjects are structured into fields and can be retrieved, rather than accessed; subject classification follows schemes, but not plans; there are LIS specialists that assign, but do not allocate, subject labels to documents to make them retrievable, rather than searchable; and that subjects are attributed, not ascribed, to a given document.

4.0 Conclusion

In this paper we described the operation of Corpógrafo as a writing assistance tool for students and researchers, using its original set of tools for terminology extraction, retrieval of collocations and identification of specialized phraseology. The examples selected for this study illustrate a small part of all the available possibilities, but they show the potential of Corpógrafo to allow writers to grasp vocabulary changes and new terms in a specialized domain. The focus of this paper is more on writing assistance as a method, rather than as a tool. Therefore, one of our aims was to demonstrate how general, freely available tools can be used by any writer to provide writing assistance, especially in LSP scenarios where a command of the terminology, phraseology and collocations is essential to write precisely and concisely. The texts included in the corpora are a research basis for extracting very recent vocabulary; hence some fundamental concepts may be less visible. The results can be enriched using a corpus of texts from a longer time interval. It is however important to note that diachronic corpora, which include texts published over a long period of time, are extremely helpful to analyze evolution in time. A systematic study is planned as future development to build solid terminology and phraseology databases in the field of KO, using the methods described.

References


Abstract:
The study reports the comparative analysis between the results of the search queries for the terms Amazônia and Amazon in Scopus and LISA databases, in the period from 2008 to 2018. Concept Theory and Domain Analysis were used in conjunction with IRaMuTeQ software in order to identify, quantify and analyse semantic distances in a sample consisting of 80 abstracts from retrieved articles.

1.0 Introduction
Inspired by the decision of ICANN - Internet Corporation for Assigned Names and Numbers, which manages the Internet address system, it was decided on 21st May 2019 that the technology company Amazon has the right to use the internet domain “amazon.com” with its variations, despite protests from South American countries that shelter the Amazon rainforest, including Brazil.

This empirical research was developed based on Hjørland's Domain Analysis (2002) and Dahlberg's Theory of Concept (1993) to investigate if the dispute between the terms Amazon and Amazônia would have any influence on the results of search expressions in the retrieval of documents in large databases as Scopus and LISA.

According to Dahlberg, “any organization of knowledge must be based on knowledge units”, that is, on concepts. Concepts are knowledge units that form the elements of knowledge systems. The concept cannot be “represented unless it is presented by knowledge units and their many possible combinations in words/terms or statements”. (Dahlberg 1993, 211).

In this sense, Hjørland argues that "it should be possible to develop the concept of the semantic distances as investigated by Brooks (1995, 1998), and to consider distances between groups and between queries and document representations". (Hjørland 2002, 445).

The present study conjugates the concept of semantic distances with the similitude analysis (graph theory) performed by IRaMuTeQ (Interface "R" for Multidimensional Analysis of Texts and Questionnaires) software applied to the terms Amazônia and Amazon as search queries.

2.0 The method
This research was based on a sample of 20 most cited articles retrieved from the same query search for Amazônia and Amazon terms in Scopus and LISA, from 2008 to 2018. The IRaMuTeQ textual corpus comprised a total of 80 analysed abstracts.

The similitude analysis results were compared based on the Domain Analysis and Theory of Concept approach. Subsequently, we also considered the applications of FAIR principles - Findable, Accessible, Interoperable, Reusable (Wilkinson et al. 2016).
Despite the existence of differences between the Scopus and the LISA databases in relation to their retrieval structure, vocabulary control, and the number of indexed journals, it was possible to analyse semantic distances assigned to the terms Amazônia and Amazon as knowledge units. The four questions indicated by Hjørland were important to conduct a qualitative-quantitative analysis in order to evaluate domains behavior and the possible transfer of meanings between them. (Hjørland 2002, 448).

To analyse terms, concepts, and keywords in the results, we used principles of the Knowledge Organization Systems (KOS), which are formed by classification systems comprising different types of relationships. (Dalhberg 1993, 212).

3.0 Results from databases
3.1 Results from Scopus

The similitude analysis of the term Amazônia unveiled that the term forest has a nuclear position, and it is formed by subclusters represented by the terms: tree, climate, amazonia, scale, pattern, datum, and specie. The central cluster is linked to the cluster represented by the term carbon. It was noticed a convergence of the subclusters to the central cluster of the term forest.

In the similitude analysis of the term Amazon, it was not possible to identify a common core term as in the Amazônia graph. It was observed that the clusters are apparently dispersed and organized in the following syntaxes: 1) the right position detaches the following terms: mechanical, turk, amazon, and show; 2) the lower position the terms: online, research, review, product, and datum; 3) the upper position the terms: mturk, datum, and system.

Figure 1: Similitude analysis between Amazônia and Amazon from Scopus

3.2 Results from LISA

The similitude analysis of the term Amazônia revealed two central clusters: a cluster identified by the term research which is composed of subclusters: brazil, brazilian, document, and digital. In addition, a cluster represented by the term information where it was noticed the presence of the term ‘amazon’ but not in a relevant position.
The similitude analysis of the term Amazon exposed a central cluster represented by the term amazon connecting the clusters identified by the terms: book, online, and mechanical, in an apparent dispersion.

![Figure 2: Similitude analysis between Amazônia and Amazon from LISA](image)

### 4.0 Discussion

In Figure 1 (Scopus), the Amazônia graph showed a closer semantic relation to the forest universe. Nevertheless, in Figure 2 (LISA), the analysis of the Amazônia graph did not apparently correspond to the syntax of the terms in the forest universe or Amazônia forest. Perhaps, because LISA is an Information Science domain database. In Figure 2 (LISA), despite the fact the term amazon is in the center, the analysis of the Amazon graph did not disclose a connection with the term forest, neither with the term Amazônia.

It was perceived in both databases that the Amazon graphs are related to the technology of the company Amazon. In this perspective, we may consider that by tagging procedure sequences in both databases, practically the same syntagmatic and lexical construction patterns are obtained.

It is worth mentioning that Scopus displayed subject areas based on its special language control in accordance with the search queries results. Also, Scopus's subject areas reflect apparently specific knowledge domains. However, those procedures are not available at LISA. Table 1 exposes the comparison of the number of documents retrieved from the search queries for the terms Amazonia and Amazon in Scopus.

<table>
<thead>
<tr>
<th>Subject Areas</th>
<th>Amazônia</th>
<th>Amazon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>12,014</td>
<td>39,031</td>
</tr>
<tr>
<td>Agriculture and Life Sciences</td>
<td>4,097</td>
<td>7,613</td>
</tr>
<tr>
<td>Environmental Science</td>
<td>2,033</td>
<td>3,757</td>
</tr>
<tr>
<td>Earth and Planetary Sciences</td>
<td>1,483</td>
<td>2,826</td>
</tr>
<tr>
<td>Computer science</td>
<td>76</td>
<td>5,759</td>
</tr>
</tbody>
</table>
The total of Amazon documents retrieved represent three times more than Amazônia ones. The distribution of documents retrieved in both terms is proportional in each subject area, except for Computer Science, which stood out in the volume of documents only in Amazon.

5.0 Considerations

The similitude analysis demonstrated that partitive, complementary or opposite syntax relations are semantically restructured to assure database retrieval procedures. In this regard, Posner's theory is relevant to discuss the term Amazon as a “comsymbol”, because the term is “independent of the recipient's context of action” (Posner 1982, 4).

According to Hjørland’s fourth assumption regarding language for special purposes, the results revealed that: “when documents are merged in databases information about implicit meanings from the prior contexts are lost” (Hjørland 2002, 446). In this sense, the results may indicate that the Amazon domain becomes dominant and merge into the Amazônia one; likewise, prior information and implicit meanings of the Amazônia cultural, social and native ethnic contexts are similarly dispersed. Furthermore, for some users, the terms Amazônia or Amazon would be the same because it is just a question of translation into the English version. For other users, the adoption of those terms may represent a strong difference regarding their value system and cultural domains (Smith 2013).

Regarding FAIR principles, the semantic distances depicted by similitudes graphs 1 and 2 pointed out that the terms Amazônia and Amazon are not interoperable, because their clustering terms distribution presented different complementary, opposition and partitive relations. Therefore, the results confirmed that “knowledge organisation systems and IR should be developed to cope with this loss of implicit information by making it explicit (database semantics)” (Hjørland 2002, 446).

This study demonstrated that the similitude analysis applied to FAIR principles is an important technical device for evaluating automatic indexing procedures in databases. The results suggested that, although we may deal with different epistemic communities with their organization of concepts and diversified genres of documents, FAIR principles are attainable goals for information retrieval systems. In addition, the findings indicate that Domain Analysis and Theory of Concept are useful analytical tools to develop databases indexing languages.

In the future, we intend to expand this method to other databases in order to improve semantic analysis in the retrieval process of terms and concepts.

Acknowledgement

This research dataset is available at the Laboratório em Rede de Humanidades Digitais (Larhud) community in the Zenodo digital repository, following Open Science practices.¹

References


¹ Dataset accessible at http://doi.org/10.5281/zenodo.3751623


A Thesaural Interface for the Basic Concepts Classification

Abstract:
We describe a thesaural interface that is being developed for the Basic Concepts Classification. This interface is particularly well-suited to the synthetic phenomenon-based approach to classification pursued by the BCC. We describe how the thesaural interface works, our plans to develop it further, and the advantages of this interface for both classifier and user.

1.0 Motivation
A classifier using the Basic Concepts Classification (BCC; Szostak 2019) would create a subject string combining terms from separate schedules of phenomena (mostly nouns), relators (mostly verbs or conjunctions), and properties (adjectives and adverbs). The resulting subject strings resemble sentence fragments. It is hoped that a classifier can move fairly directly from a key sentence in an abstract, book description, manuscript description, or object description to a BCC subject string.

Compared to classifying with an enumerative classification, the classifier is spared from having to find a complex enumerated subject heading that best fits a particular document or object. But the classifier now has to synthesize multiple terms, generally from two or three separate schedules. The BCC schedules are generally easy to navigate: hierarchies are flat and logically constructed for the most part. Yet a classifier seeking to synthesize several terms might nevertheless find it time-consuming to identify all of the necessary controlled vocabulary.

It has long been hoped, then, that a thesaural interface could be constructed that would guide classifiers to BCC controlled vocabulary. Importantly, such an interface might allow users also to enter a query in words of their choice and be guided quickly to controlled vocabulary. This in turn might encourage both public and university libraries to move back toward subject searching: Though keyword searching is easier for most library users, it is far less precise than subject searching (Hjørland 2012). A thesaural interface might potentially render subject searching as easy as keyword searching.

2.0 Design of the Interface
We are exploring the possibility that such a thesaural interface can be developed using the Universal Sentence Encoder (USE: Cer et al. 2018). One common criticism of a synthetic (post-coordinated) approach to subject classification is that a user searching for "philosophy of history" will find many documents on "history of philosophy" (Sauperl 2009) But this is only true if the search interface does not care about word order in search queries. USE does discriminate on the basis of word order, for it places each term in the context of the phrase it is embedded within.

USE is a transformer-type deep neural network, which has been trained on very large batches of text. USE can help identify synonyms for words and phrases by embedding them into vectors in 512-dimensional space. Embeddings are modeled after the idea, "you shall know a word from the company it keeps" (Firth 1957, 11) and can be seen as
a fixed length numeric representation of a text-based input. The guiding principle behind all embeddings is that if two words are often used in a similar context they likely have a similar meaning. In addition to the context of the word, USE also incorporates a token’s position in the phrase to determine its meaning. This ability to discriminate words based on their position and context is created by virtue of the way USE is trained.

During the training phase, USE consists of two principal components: an encoder sub-graph which builds a 512 dimension numeric encoding based on text input, and a decoder sub-graph which takes that numeric output as input and attempts to predict the next word in the sentence. USE maintains word order and positioning on the input phrase by adding a second 512D vector to the input which is built by overlapping different wavelengths of sin functions (e.g. sin x, 2sinx), which assign unique values to each position to a maximal length of 1024 tokens. The oscillating nature of the sin function allows the network to generalize shorter inputs to longer ones where it can potentially observe a similar distance and pattern between words used. Because of this (and other contexts observed during training where these 2 phrases are used), the phrases ‘philosophy of history’ and ‘history of philosophy’ have different embeddings.

After the network is trained, USE consists of only the encoder portion of the network, which then takes in a sentence in English, and outputs a 512D vector, as before, but now rather than predict the next word, we use the information present in that embedding to convey information about the input phrase (a sentence embedding). An interesting aspect of these embeddings which helps to add some intuition as to how they are constructed is that they can be shown to obey interesting properties when used mathematically. The classic example is that if you take the vector for the word ‘King’ and subtract the vector for ‘man’, you effectively remove all of the words associated with males from king, and you obtain the context that would surround a genderless royal (imagine words like crown, throne, rule, subjects etc.). Interestingly, now if you add the vector for ‘woman’ you will have a result which very closely matches the vector for ‘Queen’. Other common examples can be illustrated by taking a country, subtracting its capital, then adding a different capital to obtain the other country’s approximate vector (France – Paris + Rome = Italy).

Happily, USE can deal with phrases up to 1024 tokens in length, rather than just individual words. This will save classifiers and users from having to translate each word individually into controlled vocabulary. More importantly, phrases further clarify the meaning of the words they contain (for example clarifying whether "picture" is being used as noun or verb).

In order to make use of these embeddings, the entire terminology of the BCC (phenomena and relators), ISO Country and language codes, and UNSPSC1 codes used to identify goods and services within BCC have been embedded with USE (transformed into 512D vectors of floating point numbers). Further to this, an additional embedding which contains all 2 word classifications consisting of BCC relator + BCC phenomena have also been added. These embeddings are combined into a single vector array, which allows direct comparison to an unseen embedding.

When a phrase is presented to the interface, the phrase is first checked for terms which exist directly in the BCC, and is broken into sub-phrases, which will then be translated.

---

For example the phrase ‘a man dancing at a club’ is broken into three sub-phrases by the interface, around the word ‘at’ (‘a man dancing’, ‘at’, ‘a club’) which is a BCC term (NT3). This is primarily a simple heuristic which allows for breaking up input in a predictable and reasonable way. If the phrase is too long, it will likely contain too much information to be adequately translated into a 1 or 2 word BCC classification. Therefore the best results are obtained by using the most concise terminology possible with the translator.

Each sub-phrase is then embedded with USE and the resultant vectors are compared to the pre-calculated vector field of BCC (and related) embeddings. Classifiers and users can be (immediately) given ten possible BCC translations for each phrase block from which to choose. Technically, from the array the interface selects the ten nearest neighbors, using a measure of cosine similarity (cosine similarity is employed based on the assumption that vectors pointing in a similar direction have similar meaning, and ignores the magnitude of the vector).

A demonstration version of the interface can now be seen at https://sites.google.com/a/ualberta.ca/rick-szostak/research/basic-concepts-classification-web-version-2013/thesaural-interface-for-bcc. At present it deals best with shorter phrases. Readers can enter any phrase and be guided to appropriate BCC terminology (and given the BCC notation that goes along with that terminology). At present, they may have to search again if important terms are missing in the generated subject string.

3.0 Future Work

In its current state, the thesaural interface is a helpful tool. In the future as more classifications are collected as input data, a true translator could be developed that would be able to better handle ambiguity in input data without breaking the input phrase into blocks. In the prior example ‘a man dancing at a club’ the 3rd sub-phrase ‘a club’ is ambiguous without context, and the first suggestion of the translator is incorrect (a UNPC classification for clubs), and while the correct class (‘E09(901520) - nightclubs and dance halls’) is included, it was returned as a more distant match.

We are also working on algorithms that can cope with larger phrases. We can also then analyze many examples of translations. We are also working on tree structures based on the hierarchies within BCC: the translator can then appreciate that the best place to look for controlled vocabulary for a type of painting is within the category "Art" rather than "Mathematical concepts." Note that our interface can be improved over time through repeated use (and selection by users or classifiers of particular options) to better select the best BCC translation of particular queries.

4.0 Discussion

Thesauri (within information science) have almost always been developed in the past to guide users toward controlled vocabulary within enumerated classifications. The thesaural interface developed here is much better suited to a synthetic approach to classification, for it can identify the best fit by combining multiple terms within the controlled vocabulary. This thesaural interface thus reverses a potential disadvantage of a classification such as BCC. Without the thesaural interface it might be time-consuming to identify all of the controlled vocabulary necessary for a synthetic subject string (though,
again, the logical and flat structure of BCC schedulers would facilitate search for controlled vocabulary). With the thesaural interface it becomes fairly straightforward to move from a key sentence in a document or object description toward a BCC subject string. This is already the case for fairly short sentences and will hopefully become feasible in future for longer phrases. Just as it is easy for a classifier to move from a document or object description toward a subject string, it should be easy for users to move from a query in their own words toward a subject string that guides them to the document or object they seek.

Documents or objects are described in sentences. User queries are generally formulated in sentences. We have in the past attempted to translate user queries into a subject heading that is not constructed grammatically, and used that ungrammatical subject heading to attempt to identify relevant documents. It is potentially both easier and more precise to translate sentence to sentence to sentence: translate the user query into a sentence-like subject string that will guide users to documents or objects that are described by a similar sentence. The thesaural interface described in this paper can hopefully guide users and classifiers to describe a document with the same (or very similar) subject string.

Users with a precise query can thus be guided to the document(s) or object(s) they seek. Users performing exploratory searches will benefit from the fact that the interface provides them with several suggested subject strings. Users can then adjust each term in their search query to identify different sets of subject strings. They start by wondering about dogs biting mail carriers, and move on to dogs biting neighbors, dogs licking mail carriers, or cats biting mail carriers. We hope to develop a visual interface that would allow users to easily adjust their search query term by term.

5.0 Concluding Remarks

There is a dissonance in the field of knowledge organization between a body of theory that urges faceted classification and a body of practice around enumerated classification. One practical advantage of leading enumerated classifications is that they benefit from over a century of development. The thesaural interface discussed here can potentially allow a synthetic approach to classification such as the BCC to outperform enumerated classifications without the painstaking task of developing a thesaurus manually. The thesaural interface can facilitate the work of classifiers in moving from a key sentence in an object or document description toward a BCC subject string. It can so facilitate user queries that these can be as easy as keyword queries – but provide much more precise results.

References


The Role of Knowledge Organization in Scientific Communication
An Overview on JCR's Psychology Journals Guidelines about Title, Abstract and Keywords

Abstract:
The aim of this exploratory study on writing the title, abstract and keywords of articles was to verify and analyze guidelines or strict rules for authors in scientific journals in the field of Psychology. A combination of quantitative and qualitative techniques has been used in a representative sample of journals in Psychology indexed in the Journal Citation Reports (JCR), considering the first eight titles of each of the four quartiles. The sample consisted of 32 journals out of 77 (41.5%). A previously adapted and tested spreadsheet has been used. The data collection and processing of 38 items were grouped into categories related to identification data, guidelines for the title, abstract and keywords of the scientific article. The results obtained reveal that the 32 journal titles belong to 11 different editors; all of them have some kind of style manual with orientations that vary between deep and more superficial; and most of them have guidelines for writing the title (75%, n. 24), abstract (96.8%, n. 31) and keywords (71.8% n. 23). The use of a graphic summary and a video summary stand out as new ways of disseminating the content of scientific articles. It is concluded that these aspects are aligned to a greater or lesser extent with the guidelines of the ISO Standard and the APA Manual. However, they are more related to the protocols of editorial management procedures for the dissemination of scientific work in a digital environment than to the intrinsic quality of documentary products associated with the organization and representation of science.

1.0 The aim and scope of the study
The broadest sense of knowledge organization (Hjørland 2008) is about how knowledge is organized socially. It includes the production and dissemination of knowledge as a social division of mental work. It means that scientific journals, in a broad sense, are the result of the social organization of knowledge production and dissemination. In a proposal for a model of scientific publications as a process of organizing knowledge in a broader sense, Nahotko (2014), presents the relationships between organization of data, information and knowledge in the research cycle that results in specific organizational structure to resources that facilitate the user access to knowledge. In this sense, in addition to the text that presents language and textual structure characteristic of a scientific article, separate entities, such as the title, abstract and keywords are representations of knowledge. Actually, in the formal writing of articles, the title, abstract and keywords have been essential elements of knowledge representation and organization providing intellectual access to catalogs, bibliographies, indexes and databases (Hjørland 2003). These elements are currently essential to promote access to publications, as they are the main points of access to documents. Indeed, it is on them that users focus their research to retrieve documents and it is also based on them that they decide whether to consult the full text (Sabadini, Sampaio, and Koller 2009).

The international standardization that mentions the elements “Title, Abstract and Keywords” in its guidelines is the Standard ISO 215: 1986, entitled “Documentation -
Presentation of contributions to periodicals and other serials”, still in force and current according to information provided by ISO. According to Aparício, Guillermo, and Liberator (2016), the Standard ISO 215: 1986 addresses the Title, Summary and Keywords for the presentation of scientific articles. It provides guidance so that the title is brief, concise and clear regarding the most relevant aspects of the work without including non-significant terms such as "research", "study" or "formulas", as well as abbreviations, acronyms or codes that are not in common use. As for the abstract, it determines that it should have between 150 and 200 words and that it should contain, in a single paragraph, the topic of the work, the methodology and conclusions without including interpretations or criticisms, bibliographic citations or referring to figures or images of the text. It also recommends the inclusion of about four to six keywords after the abstract. For the assignment of keywords, it indicates the use of thesaurus.

The Psychology field, however, has developed since 1929 a style and standardization manual (American Psychological Association 2010), The Publication Manual of the American Psychological Association, currently in the 7th edition (American Psychological Association 2019). Specialized in academic, print or electronic publications in the field of Psychology, it also offers style guidelines and standardization of article writing. With regard to the title, the APA Manual (American Psychological Association 2010) draws attention to its strategic function of serving the indexing and compilation in countless reference works including web browsers. The guidelines indicate that the title is self-explanatory and words with no purpose such as "method", "study", "research", "results" should not be used to avoid confusing the indexers. It also advises that the title does not exceed 12 words without the use of abbreviations and acronyms. For the abstract, the guidance on the representation function for indexing and retrieval in reference works is repeated, with four quality requirements, be precise, concise, coherent and legible and not be evaluative, in addition to writing guidelines for 5 different types of articles: empirical study, literature review or meta-analysis, theoretical article, methodological article and empirical study.

As for the number of words, the limits can generally vary between 150 to 250 words, however, the orientation is for the authors to respect the limits established by the journals. The keywords accompany the abstract and must serve to guarantee the location of the article by the user.

Fujita, Agustín-Lacruz and Terra (2018a; 2018b), carried out an exploratory study to observe and analyze the guidelines provided to authors on the writing of the title, summary and keywords of their articles in a representative sample of the Librarianship and Information Science (LIS) journals and Communication Science (CS) indexed in the Journal Citation Reports (JCR). The results revealed a tendency to standardize general indications and criteria and little instruction on titles, abstracts and keywords more related to editorial management than associated with the organization and representation of knowledge. In view of this background, it is pertinent to find out to what extent scientific journals in the field of Psychology define guidelines or strict rules for the authors to follow when submitting text proposals for publication.

2.0 The Method

In 2018, JCR included 77 journals from Psychology field. For data collection, a sample was created choosing the eight first journals of each one of the four quartiles. In that
respect, the sample consists of 32 journals, representing 41.5% of all the JCR 2018 Psychology titles. In order to organize data collection, an adapted spreadsheet previously tested and used (Fujita, Agustín-Lacruz, and Terra 2018a; 2018b), was applied. In December 2019 and January 2020, the website of each selected journal was analyzed in order to collect data about the 38 fields of the spreadsheet. The data collected include: generic information about the journals (title, publisher, ISSN, position ranking, impact factor, web address), guidelines for providing article titles (writing, translation, number and types of words, type of information content, subtitle, use of abbreviations or symbols and style), guidelines for abstracts writing (translation, abstracts differentiation according to article category, number of words, structure, style, use of abbreviations or symbols and inclusion of bibliographical references) and guidelines to provide keywords (translation, number of keywords, separation punctuation, vocabulary control and type). Recommendations and rules for the authors who want to publish in scientific journals are now available on the websites of each journal. Thus, data collection on journals websites seems to be an adequate strategy to create a knowledge base on guidelines or rules about title, abstract and keywords that journal editors and publishers defined for each publication. Data were analyzed from a quantitative point of view and subsequently subjected to a more in-depth analysis, focusing on the analysis of the website content in relation to specific aspects of each criterion used.

3.0 The Results

The main results about collected data will be presented and analyzed next, considering quantitative and qualitative approaches.

The publisher with most titles in the sample is Taylor & Francis (11 titles, 34.3% of the sample). American Psychological Association and Wiley have both four titles (12.5% of the sample). Elsevier (three titles) and Sage (three titles) are also represented, as commercial publishers. Annual Reviews, a non-profit publisher, counts two titles. The sample includes five more publishers (commercial and non-profit) with one title journal each one.

All the journals present some kind of style manual, giving deep or more superficial guidelines for authors to write their articles.

Titles - Within these style manuals, a vast majority of them (75%, n. 24) presents guidelines for writing titles. The number of words for titles is defined for 31.1% (n.10) of the sample with recommendations like “no more than 12 words”, “title should be around 150 characters”, “short title of not more than 45 characters (including spaces)”, “25 words maximum” or “title of less than 40 characters”. There is also a journal with a more generic guideline as “do not use long article titles”. Only three journals (9.3%) include tips about the kind of words to be used in the title and they all refer to include a few keywords in the article. This recommendation is related to Search Engine Optimization (SEO), and some journals have another section related to this matter recommending authors to include a few keywords of an article in its title and to not use long article titles. With regard to the information content of the title, four journals (12.5%) give some kind of guideline, namely “there are three rules when it comes to deciding on your title: make it concise, accurate, and informative” or “the title should be concise, omitting terms that are implicit and, where possible, it should be a statement of the main result or conclusion presented in the manuscript”. Avoiding abbreviations and acronyms is
recommended by three journals (9.3%) and this could be also related to SEO optimization. Titles that are a mere question without an answer, unambitious titles (like the ones starting with "Towards", "A description of", "A characterization of", "Preliminary study on"), vague titles (as starting with "Role of...", "Link between...", "Effect of...") should also be avoided.

Abstract - Almost all journals (96.8%, n. 31) present guidelines for abstract writing. They are included in a specific topic of the style manual or, in one case, only available in the information system management of the journal for submitting articles. When included in the style manual, rules for abstract are included in topics with very different names such as “Article Components | Required elements”, “Checklist for Manuscript Submission | Title Page and Abstract”, “Checklist: What to Include”. In two cases (6.2%), there is a specific topic entitled “Abstract”. Merging guidelines about title and abstract in the same topic is common in the sample. With regards to the number of words in the abstract, 28 journals (87.5%) set some kind of rule. Abstracts can have between 100 and 300 words depending on the journal. The number of words can vary within a range (e.g. “the abstract must have between 150 and 250 words”) or even have an exact number (e.g. “do not exceed 300 words”) with no minimum number. Some journals define different rules for different kinds of articles, like “no longer than 250 words for regular articles, and no longer than 150 words for clinical notes”. Nine journals (28.1%) do not include any reference to abstract structure. However, structured abstracts are used in 15 journals (46.8%). A structure including Background, Methods, Results and Conclusions is required in six (18.3%) journals, other journals require other kind of structure with little differences between them. It is also important to underline that five (15.5%) journals recommend graphical abstracts and two (6.2%) recommend video abstracts. This is a new approach to abstracts encouraging new ways to disseminate scientific work as one journal states that “the graphical abstract should summarize the contents of the article in a concise, pictorial way, designed to capture the attention of a large number of readers”.

Keywords - A vast majority of the journal sample (n. 23, 71.8%) present guidelines about keywords. Usually, there is not a specific topic about keywords in the style manual or it is even necessary to register in the submission management system to know how many keywords to include. Some journals also accept short phrases and do not require keywords. The number of keywords can vary from one to 10 keywords depending on the journal. Only three (9.3%) journals ask for semicolons to separate keywords. Seven journals (21.8%) use some kind of controlled vocabulary. Two of them use the NIH MeSH terms, one uses Index Medicus and four provide their own list of keywords.

4.0 Conclusions

We believe that the text includes a concise and systematic study based on observation, which has value in the current context; marked by the availability of a large amount of information which can be quickly accessed online, and is assisted by Search Engine Optimization (SEO) in discovery systems. It is concluded that these aspects are aligned to a greater or lesser degree with the guidelines of the ISO Standard and the APA Manual. However, they are more related to the protocols of editorial management procedures for the dissemination of scientific work in a digital environment than to the intrinsic quality of the documentary products, associated with the organization and representation
of science. Abstract and keywords are documentary products resulting from processes of organizing knowledge in a more specific sense, such as indexing and abstracts preparation. Both have advanced theoretical and methodological development in the field of knowledge organization whose foundations could, in fact, be applied to the improvement of scientific articles. It is observed, as an improvement factor, that some journals use some type of controlled vocabulary to define keywords.

On the other hand, in a broader sense, it is observed that the organization of knowledge is present in the guidelines and recommendations of style in the field of Psychology, strongly influenced by the APA Manual. They are essential to characterize the textual structure of scientific articles as a social organization of knowledge in Psychology. Expanding and deepening new applied research on the processes of organization and representation of knowledge in the editorial policies of journals can be an interesting approach to the field. Proposals for standardization and for the authors' guidance manuals for scientific articles are also relevant.

References


The Social Role of Knowledge Organization in Evidence Based Health

Abstract:
Evidence based health is a form of knowledge production that aims to provide the best information for decision-making. The knowledge produced in Evidence based health, such as systematic reviews and technical-scientific opinions should be structured in such a way that clarifies the steps taken in the search for information and its retrieval, thus allowing access to the studies, which can be done by adopting knowledge organization principles, instruments, and methods. Through planning and systematization of information search and retrieval processes, knowledge organization has fulfilled its social role inasmuch as it has helped develop better approaches to patient care by managing health resources. As a result, this study presents an information search and retrieval methodology to build best practices within health technology assessment in the context of Evidence Based Health care.

1.0 Introduction
The goal of this study was to highlight the social role of knowledge organization (KO) in Evidence Based Health (EBH). The meaning of the term “social” here refers to peoples’ well-being and cooperation for a good life in society. The use of mapping among knowledge organization systems (KOS) can contribute to peoples’ well-being by being part of the production of scientific knowledge, whose purpose is to offer the best health technologies, according to patients’ needs and available resources. The role of KOS in this scenario is to provide a larger sample of terms and synonyms derived from the vocabulary of various discursive communities to develop sensitive search strategies, allowing for retrieval with higher recall and relevance.

The main question that drives this analysis was: can the planned and systematized information search and retrieval process be considered part of the evidence in the context of EBH? This study presents a proposal for using KOS mapping when building sensitive search strategies for searching and retrieving health data. The goal was to systematize and rationalize the search process to support healthcare management and care, especially in decision-making processes in the context of health technology assessment (HTA). This assessment is based on the principles of quality and excellence of the connections among research, policy and management in all its phases (incorporation, diffusion, abandonment), in due time (Brazil 2013).

The health technology assessment process, which is part of EBH, provides healthcare managers with support to make coherent and rational decisions about the inclusion of health technologies (Brazil 2011). Health technology assessment contributes to the well-being of people by enabling the provision of health technologies that are appropriate to the needs of patients. Questions regarding the best medicine, surgical technique, or prosthesis for a group with particular characteristics such as age, sex, place of residence, and living conditions and habits are some of the decisions made routinely in hospitals, and these directly interfere with the quality of life of thousands of people.

Systematic reviews (SR) and technical-scientific opinions (TSO) support HTA. Systematic reviews and TSOs are examples of evidence that presents knowledge that can
enhance the decision-making process, making it adequate to resource management and the well-being of patients. SRs are a “method of scientific investigation that uses the planning and gathering of original studies, synthesizing the results of multiple primary investigations” (Universidade Federal de Santa Catarina 2010, 27). An essential part of planning an SR is carrying out a literature review (Brazil 2012). In contrast, TSOs do not require such an extensive and comprehensive literature review and are quicker to prepare, representing a systematic and comprehensive account of the knowledge that can be found (Brazil 2011).

The structure of SRs and TSOs lies in the presentation of the information search and retrieval process (Brazil 2011; 2014); thus, they can be considered part of the process of construction of evidence because one of its elements is its reproducibility. The evidence describes the research problem, the consulted sources and KOs, and the sensitive search strategies used in each source, with due justification. The methodology proposed by Andrade (2015) and Andrade and Lara (2019) highlights the role of information exchange between specialists in the field of health and KO in defining and understanding research problems and mapping terms and their synonyms while using the Library Medicine’s Metathesaurus. The search, retrieval, systematization, and presentation of search results to healthcare experts are also included in the referred methodology. The search strategy model makes it possible to retrace the steps of the search and identify changes in it.

The social role of KO in EBH is the use of the information search and retrieval methodology with mappings among knowledge organization systems when creating sensitive search strategies. These offer a greater number of synonymous terms, from different thesauri, terminologies, bibliographic classifications, billing lists for health services, which were built according to local needs and the languages of specialties from different discursive communities.

2.0 Methodology

This was an exploratory and bibliographic study. We adopted the theoretical-methodological approaches of the organization and retrieval of information and knowledge (Hjørland 2008; Lara 2013) and of evidence based healthcare (Atallah 2004; Universidade Federal de Santa Catarina 2010; Brazil 2011; 2012; 2014). Two experiments were conducted: a) a case study in orthopedics and traumatology and b) an evaluation of the National Library Medicine Metathesaurus that enables the simultaneous search of terms in several KO systems (Andrade 2015). The case study aimed to build generalizations about the identification of research problems, the search and retrieval of information from databases, focusing on mapping terms and concepts to develop search strategies. To assess the metathesaurus, the researchers tried to identify the ways how the metathesaurus functions, its use and results. For this analysis, researchers used the ISO 25964-1/2:2011, Information and documentation – Thesauri and interoperability with other vocabularies, which offers recommendations to identify and classify equivalences between terms (International Standard Organization 2011a; 2011b).
References


Social Network Communication and Effects on Innovation
The Case of the Agrifood Sector in Algeria

Abstract:
We studied Algerian researchers in the field of agronomic research to analyze their information seeking behavior and evaluate their level of interest in sharing knowledge. “Communication… is one of the core requirements for carrying out collaboration, or maintaining any kind of productive relationship” (Shah 2010). Therefore we propose a timely framework for sharing knowledge to increase collective intelligence and spur innovation to rise agricultural and food production in Algeria.

1.0 Introduction
The paper is concerned with innovation and technology diffusion in the Agrifood Sector in Algeria. Our design uses a Knowledge Organization System that makes connections between research results on a given concept and innovations for which such results may be relevant. In Algeria the dysfunction in communication impedes companies and research institutions due to the absence of information system and a lack of culture of sharing information. Information and communication technologies, especially such platforms, make it possible to exploit collaborative processes that underlie collective innovation (Iskia 2011). For Rogers and Schoemaker (1971), innovation can take the form of an idea, a practice or an artifact material, each of which has an attribute of novelty, whether tangible or intangible. If the common sense given to innovation is based on the notion of novelty (a change of state of an element from the previous state), the difference is often presented in the literature as positive and associated with an idea of improvement. Boldrini (2005) indicates that there is no innovation without sanction by the market, no innovation without design and no innovation without an innovative company. This brief synthesis presents some aspects of the dominant theoretical understanding of innovation and attempts to draw some tracks that would be useful to guide thinking about an information management system.

2.0 The method
We carried out two surveys, using a mix of quantitative and qualitative methods.
Survey 1 asked researchers specialized in agronomic sciences about their information practices through collaborative tools. In September 2018 we distributed 250 questionnaires to researchers at Institution of Agricultural Research and Higher Education in Algeria. By November 2018 we received 154 returns (62%).
Survey 2 asked practitioners working in Agrifood companies that are part of the “Cevital Group” (we did not study small farmers) to understand how intellectual capital is transferred from research to companies and how it sparks innovation. In October 2018 we distribute 250 questionnaires. By December 2018 we received 140 returns (56%). To gain deeper insight into the effect of communication on innovation, we are carrying out follow-up interviews.
3.0 Results

3.1 Communication, technology transfer and collaboration for researchers

We examine the strengths and the weaknesses for researchers.
- Motivation of researchers in sharing research results encourage an execution of the research policy to create a strong synergy between research programs and the economy sector.
- Researchers are generally interested in collaboration by social media platforms (General social networks, Academic social networking sites, Discussion groups).
- Information seeking behavior for Algerian researchers is not very open to share knowledge using collaborative tools: Listserv groups, blog, e-mail, local database.

The interview shows the organization's integration in the transfer of technology of university research to AgriFood sector. The researchers attest that the types of intervention with companies are more specifically focus on technical reports, scientific articles and opinions. They state that the access to information through social networks make more innovative. Some of the researchers participate in the development of process innovations in companies thanks to their workshops.

3.2 Management and sharing knowledge for practitioners

We quote some strengths and weaknesses for practitioners.
- Availability of information in the company using both formal and informal sources.
- Cooperation shows better performance and good productivity for the company.
- Collaboration by social media permits a synergy: research-innovation-development.
- Lack of social media collaboration with researchers using academia social networks.

Concerning the interview for transfer of technology, the research of information on technologies carried out by expertise. The information that led to innovation is precisely agreements and reports. To improve technologies and product models practitioners focus on control and quality.

3.3 Proposition of social media tools for researchers and practitioners

- Dissemination of knowledge using blogging tools (Blogger, WordPress, Posterous) can provide to researchers to post their papers to have visibility.
- Collaborative writing tools by Google Drive (Google Docs and Dropbox).
- Social citation-sharing to manage references’ researchers and to share their scientific production as Mendeley and Zotero.

4.0 Conclusion

The survey showed reluctance of information within and outside the company. Using blogs liberates practitioners to share knowledge with their colleagues and to use the research results from scientists for innovation. The establishment of social media platform in AgriFood Sector in Algeria will emerge the aspect of collective intelligence in order to create synergies between national research programs, and between the research and the economy sector.
References
Organization and Sharing of Knowledge on Selective Household Waste Collection for Hygiene and Sanitation in the City of Yaoundé, Cameroon

Abstract:
With reference to Michel Beau who conceives knowledge organization as the ability to model a documentary memory that functions as a physical extension of our individual memory, thus allowing the dynamic sharing of a space for knowledge capitalization (knowledge base) and knowledge transmission, both reactive and prospective, our poster explains the use of knowledge organization models in social marketing and communication for development. The target audience here is community, the inhabitants of a city. In other words, the aim was to decipher the extent to which this model is applied in the management of local authority in order to encourage the adoption of behaviour deemed to be favourable to the general interest. This study focuses especially on the organization and sharing of knowledge on the selective collection of household waste to the populations of the city of Yaoundé in Cameroon, with a view to promoting the recycling and recovery of waste by category. In this sense, we have carried out a critical analysis of the means and mechanisms deployed by the actors of hygiene and sanitation of this city, to organize and make available to the public, the knowledge and know-how related to the sorting of garbage. Interviews with resource persons in the field of household waste management and sorting within the organisations concerned in the city of Yaoundé, the use of their documentation, participant observation and an opinion poll of the inhabitants of Yaoundé enabled us to trace the circuit of waste collection and management in Yaoundé on the one hand, and that of knowledge in the field of waste sorting on the other hand. It emerges that, despite its impact on health of the populations, knowledge on waste sorting is disseminated in the city of Yaoundé in a marginal way, precisely according to the capitalist interests of the actors; while the knowledge needs of the populations on the issues are widely expressed. In addition to improving the system for collection of recyclable waste, a digital application on waste sorting in Yaoundé could help raise the awareness of its inhabitants by answering precisely the questions they have on this topic.

1.0 Research question
How is knowledge on household waste sorting organised and shared with the inhabitants of the city of Yaoundé to promote their effective involvement at this level of waste management? How could the principles of knowledge organisation help meet the knowledge needs of the inhabitants of Yaoundé in this respect?

2.0 Motivation of the research in relation to knowledge organization and the conference theme
The use of knowledge organisation models in social marketing and communication for development to promote the adoption of behaviour of general interest.

3.0 Methodology
Interviews were conducted with resource persons in the field of household waste management and sorting in the following institutions: Yaoundé Urban Community, HYSACAM, Namé Recycling, Solidarité technologique, Era Cameroon.
Documents received from these institutions permitted the constitution of a research corpus.
The study population considered for the survey was made up of 150 persons living in the 7 district areas of the city of Yaounde.
Participant observation was conducted at the Hysacam processing plant (Nkolfoulou) and with the local population.

4.0 Major findings

In the city of Yaoundé in Cameroon, household waste collection is the subject of a contract between the Yaoundé Urban Community and Hysacam (Hygiene and Sanitation of Cameroon). The latter is responsible for collecting all the waste produced by households in the city and disposing it in its dumping zone.

According to epidemiological data from the Ministry of Public Health, in 2019, hygiene-related diseases are the top five reasons for consultation and hospitalization in all regions of the country. The urban environment is essentially unhealthy, and poor management of solid, liquid, gaseous and radioactive waste remains a major public health concern throughout the country.

Figure 1: River in the valley between the hills of Mvolyé and Ngoekélle, Yaoundé 3, taken on 22 December 2019 at 16h

One of the most visible consequences of the lack of waste sorting in the city of Yaoundé is the overloading of waterways with plastic bottles. This favors the stagnation of water and the proliferation of mosquitoes which are malaria vectors.

According to our survey, 71% of our sample do not sort household waste due to a lack of interest or real motivation. They seem to express a need for knowledge about the methodology for sorting their household waste (what to sort? and what to do with the sorted waste?). Moreover, although they are aware of the importance of sorting waste against environmental pollution, they seem to need incentives and facilitation measures initiated by the competent authorities in this area, such as the provision of selective garbage bins.

The civil society organizations involved in hygiene and sanitation are the only ones to bring to households the knowledge on the sorting of household waste. This is limited to the pragmatic aspects: separate them and we will take them off your hands. In addition, the channels for disseminating this knowledge are limited to specific areas or neighbourhoods. Hence the low rate of practice of waste sorting (29% of our sample).

Sorting household waste in Yaoundé is essentially a commercial activity, and the knowledge shared by the actors involved is mainly about the monetary value of the
waste. Waste that is bought back cheaper or not bought back at all on the Cameroonian waste market is not sorted. They are buried or abandoned in gutters and then end up in rivers, regardless of the negative consequences for the environment and the health of the inhabitants. This is precisely the case for plastic bottles and packaging.

In addition to the implementation of a model of household waste management that effectively involves the population in the practice of sorting, the needs of the inhabitants of Yaounde for knowledge on the interest of sorting both for the environment and for their health are important.

5. Conclusion

No principle of knowledge organisation seems to be applied for the time being in the framework of this awareness raising exercise. However, it would be beneficial to model low cost digital knowledge organisation tool (Paraponaris and Simoni 2006), that should allow the Ministry of Public Health, in partnership with the Yaoundé urban community, to share with the inhabitants of Yaoundé, in a dynamic way, knowledge that should answer their specific questions which include:

1. Why sort waste (individual and collective benefit)?
2. How do you sort waste?
3. What to do with the sorted waste?
4. What happens to the waste once it has been sorted?
5. What are the consequences of unsorted waste on my health and my environment?

References
Devising a Concept of User for Archival Science
An Analysis of the Brazilian Scientific Literature

Abstract:
This poster presents an overview of Archival Science literature on various concepts of use and users of information in the Brazilian scenario. Its purpose is to map scientific articles related to uses and users of archives, and find variations of applied terminology. We plan to show the extent and influence of the production on users of information in Brazilian Archival Science. In this sense, we implemented a quantitative analysis of key-words related to users in a Brazilian scientific database. The initial findings present a sketch of different concepts employed to address users and their information requests. The research perspective suggests the application of knowledge organization techniques and methods to devise concepts of user of information.

1.0 Introduction
It is well known that the aim of archives is granting access to information to fulfill the requests of their users. To better accomplish that, it is critical to understand the user of information in archives. Thus, we started an investigation of scientific articles and concepts adopted in Brazilian Archival Science on the subject.

Here are shown the preliminary results of an analysis of a renowned Brazilian Information Science database called BRAPCI, which contains indexed articles from 57 scientific journals since 1972. There is a particular interest in the literature on Archival Sciences, and how the notion of user of information in archives is conceptualized in Brazil. The theoretical assumption is based on the need to organize knowledge on concepts related to the users of information. The starting point is the premise of the role and power of representation in Western culture, from Foucault (2000), and the unfolding of a philosophy of knowledge organization (KO) present in Hjørland (2008; 2014) and Cooper and Mazzochi (2011). Therein, it is fair to question the extent to which the philosophical construction in KO can contribute to the conceptualization of user studies and information users in archives.

2.0 Methodology and results
The initial stage was the calculation of occurrences of terms regularly used in academic articles to deal with the users of information and their research trends. The terms used for data collection in this research were: Use/Usage; User; Use and User; Need; Demand; Literacy; and Behavior. The examination took place according to the database parameters and searched through the titles, key-words and abstracts of the indexed articles. We first scrutinized the expressions on their own, then in association with the term Archive/Archival. The number of papers containing these terms in general was found, together with the amount of articles where the words appear associated with archives. The primary intent was to identify the number of concept occurrences and the

1 Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (Brapci).
https://www.brapci.inf.br/
nuances of terminology, in order to make future inferences on the conceptual framework for user studies in Brazilian Archival Science.

The results show term occurrences on their own (SINGLE) and associated with archives (COMBINED). There has been a further review of articles on combined occurrences to sort out those that consisted of suitable user-related themes. We opted to discard those without the analyzed concepts. This happened because most words are widespread, so they could be present in texts, but not directly related to the thoughts as examined here. These refined results are shown in the SPECIFIC column. Table 1 presents the three categories:

<table>
<thead>
<tr>
<th>TERMS</th>
<th>SPECIFIC</th>
<th>COMBINED</th>
<th>SINGLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>USER</td>
<td>26</td>
<td>282</td>
<td>3513</td>
</tr>
<tr>
<td>USE/USAGE</td>
<td>10</td>
<td>344</td>
<td>4384</td>
</tr>
<tr>
<td>NEED</td>
<td>9</td>
<td>358</td>
<td>2818</td>
</tr>
<tr>
<td>DEMAND</td>
<td>8</td>
<td>102</td>
<td>1085</td>
</tr>
<tr>
<td>USE AND USER</td>
<td>5</td>
<td>66</td>
<td>834</td>
</tr>
<tr>
<td>BEHAVIOR</td>
<td>3</td>
<td>44</td>
<td>1003</td>
</tr>
<tr>
<td>LITERACY</td>
<td>1</td>
<td>141</td>
<td>2182</td>
</tr>
</tbody>
</table>

3.0 Evaluation of the findings and conclusion

The articles containing these concepts in Archival Science are a small part of the universe of the Information Science academic production. The specific results show that the theme users has feeble relevance for the archives, as stated in both Brazilian (Jardim and Fonseca 2004; Oliveira 2004; Gasque and Costa 2010) and foreign (Tibbo 2002; Rhee 2015) literature.

The results of the SPECIFIC column represent only 27 single articles with themes related to archives. The articles date past 2004, showing that it is a new subject of interest. Out of the 27 articles on user-related issues from 2004 to 2019, almost all (26 articles) included the term “user”. The terms “use”, “need” and “demand” are all present in about a third part of the articles. “Literacy” and “behavior” are in the last positions, showing a modest use within Archival Science.

Preliminary quantitative data reveals another problematic issue for the study of users of information: the absence of a meta-informational production capable of ordering and supporting the discussion about the concept of user in different contexts. The data indicates the convenience of a strict methodology for organizing knowledge to support archival institutions in understanding users and their criteria for searching and using information. So far, “user-oriented tradition” has “almost totally neglected epistemological theories and thus confused the concept of ‘users’ and the concept of ‘subjectivity’” (Hjørland 2008, 93-94).

The archives’ users should be further explored in academic research to understand what society expects from archives and allow users to contribute to archival activities. Granting access to information can only be fully achieved by corresponding in some way to the needs of users. In the Brazilian scenario, archivists have yet to learn to love their users (Tibbo 2002).

Further studies of international production would bring a better understanding of the bigger picture. Nevertheless, in Brazil, Archival Science shows a lack of epistemological reflection on users of information and their specific requests. Essential concepts such
as Information Behavior and Information Literacy are close to nonexistent. So it is feasible to conclude by stating that the theme of users of information in archives entails a more solid theoretical development to support the progress of such a crucial area of study.

References


Abstract:
The present research intended to analyse what is the perception from International Society for Knowledge Organization (ISKO) about Ethical studies in the field. To do that, was applied a discourse analysis in the papers with social contents presented in the international events of the society. Was possible to see a criticism to misrepresentation in languages considered universal, the exclusion of marginalized groups in society, and the change of paradigms of information access from the technological development.

1.0 The aim and scope of the study
Considering the relevance of ethical studies in knowledge organization, facing an increasingly globalized and multiple world, this study intended to analyze how the International Society for Knowledge Organization, while an important representative in knowledge organization studies, sees the Ethical studies in the field.

The study aimed to produce a theoretical framework about the main researches in this theme and reveal what is their point of view and ideas not declared expressed in their articles.

2.0 Methodology
For methodological aspects, we used the discourse analysis, arising from Pechêux studies and related to Foucault texts. This methodology aims to highlight the ruptures of ideologies and regards dialectical relations between dominated and dominant. The analysis was made from a search in ISKO international events papers, using the terms ethic*, gender, privacy, culture, prejudice, and bias, presents in the titles and abstracts of the articles. 72 articles were retrieved and analyzed using the Sketch Engine qualitative analysis software. This tool allows explore how language works, from the analysis of terms present in the text and the forms that they were found - as nouns, the verbs that accompany them, as an object, or as a subject, among other aspects. To discourse analysis, we also based on works by Barros (2012; 2017).

3.0 Results
Regarding the radical "ethic*," was observed that the highest incidence of a modifier of the term is "transcultural," primarily due an article by García-Gutiérrez (2002), who proposed the use of transcultural ethics for KO. The author states that a culture of border is necessary, in which rigid systems are rejected, approaching and accepting the perspective of the permanent mutation of individuals and societies. In this sense, the different cultures are accepted by the representation scheme, which is configured openly and horizontally, in contrast to the traditional systems, viewed hierarchically.

The relationship cloud of the term "gender" was more dispersed. One of the criticisms found is about the dichotomous perspectives and binary oppositions. In repre-
sentation languages, these conceptions are adopted: either belonging to one class or to another, usually in opposition. However, the discourse that legitimates this binary view is detrimental to those who find themselves in the middle between two notions, who do not feel represented by either. Besides, the opposition view itself is harmful, since it puts the terms in an antagonistic way, reinforcing the difference between the concepts. In this sense, those who escape legitimized rules by systems are once again excluded, considered as abnormal, wrong, or not belonging to the group.

The concept of prejudice is discreetly addressed in the articles. The emphasis is placed on the connection between the term analyzed with the concepts of bias and preconception, the first concerning the biases that can be found in the languages of representation and the second, the preconceived ideas on a given subject, representing that group in a pejorative way. In the analyzed documents, these ideas are worked in line with issues such as justice in the treatment of people or with the prejudices that can be spread on social networks, which justifies the highlighted presence of the term with technology.

The last term analyzed and which has a strong relationship with the term prejudice, is the concept of bias. It was noticed that the term bias has a strong relationship with the first term analyzed, related to cultural aspects. These relationships can be identified in investigations about the concepts of cultural hospitality and warrant in representation languages inserted in multicultural contexts and in how classifications can lead to a culturally biased view of knowledge, even those considered universal, can be considered harmful in specific contexts.

4.0 Conclusions

From this analysis, it was possible to identify that these concepts have had their presence since the first congress and that the main vision of the society is the criticism to marginalized groups, and new perspectives from the technological context. More specifically, the investigations evidenced how harmful the use of tools considered universal to specific contexts could be, since the construction of these tools starts from an anglo-Saxon view, excluding aspects that escape this so-called normality, marginalizing individuals considered as "other."

References


Educational Practices of Knowledge Organization in Iran  
A Historical Review

Abstract:
In this paper, we discussed the educational practices of the Knowledge Organization in Iran. This paper is a historical review. In order to analyze the educational practices of Knowledge Organization related researches are reviewed. Also, curriculums of The Iranian Ministry of Science, Research and Technology (MSRT) are analyzed to identify courses related to Knowledge Organization and their units. In past years, some courses have been modified or revised to meet user needs.

1.0 Introduction
Knowledge organization is one of the most important courses in LIS education. Core elements of this field have a significant impact on developing curriculums of LIS (Pat-tuelli 2010; Alajmi and Rehman 2016). Therefore, it is necessary to analyze the last curriculums of educational levels to identify the compatibility degree between modern Knowledge organization approaches with current Iranian curriculums. In 1965, the first library department in Iran established at the University of Tehran. Also, standard practices of cataloging and classification were started in 1968 by the establishment of TEBROC (Tehran Book Processing Centre) (Fattahi 2007). Later, in 1983 TEBROC was incorporated into the National Library of Iran. These two agencies had an important role in the standardization of cataloging in Iran. This study aims to investigate new trends and developments of KO education in Iran. In this study we addressed 3 main questions. These questions are:

- What developments are considered in Iran related to education in KO?
- How Iranian LIS curriculums are changed based on new developments in KO?

This study is a historical review. To gather the required data curriculums of The Iranian Ministry of Science, Research and Technology (MSRT) and The Iranian Ministry of Health and Medical Education (MHME) and related researches were analyzed.

2.0 Results
Anglo-American cataloging Rules second edition (AACR2) and MARC format are still the most crucial technologies in Iran. In 1992 Rahmatollah Fattahi translated AACR2 into the Persian language (Fattahi 1992). On the other hand, the National Library and Archives of Iran (NLAI) developed the IRANMARC format based on UNIMARC. Library of Congress Classification (LCC) and Dewey Decimal Classification (DDC) are widely used in Iranian Libraries and they are included in the curriculum of Information Science and Knowledge. Also, in medical universities NLM classification and Medical Subject Heading (MESH) are taught to students. Likewise, in many universities of Iran, laboratories of Knowledge Organization are enriched to meet students’ needs. For instance, Shahid Beheshti University has launched liblab. Iranian Authority files, thesauruses, classification schedules are prepared in these laboratories.
<table>
<thead>
<tr>
<th>Degrees</th>
<th>Courses related to KO</th>
<th>Theory</th>
<th>Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undergraduate</td>
<td>Descriptive and analytical cataloging</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Dewey Decimal Classification</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Library of Congress Classification</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Designing library software</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Computerized organization of references</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Indexing and abstracting</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Designing digital library</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Master (Information Management)</td>
<td>Information Representation</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Information Storage and Retrieval</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Master (digital library management)</td>
<td>Digital resources organization</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Digital libraries user interface</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Doctoral (Information and Knowledge Retrieval)</td>
<td>Natural language processing</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Semantic web and ontology</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Information architecture</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Text analysis and knowledge mining</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Doctoral (Information and Knowledge management)</td>
<td>Knowledge audit and Organization</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

More than 100 books have been written in this field to cover various aspects of Knowledge organization (Malek Mohammadi and Haji Zeinolabedini 2017). Fattahi and Pazooki (2018) showed that 25 M.A. theses and 7 Ph.D. theses have been written in the new approaches of Knowledge Organization. As stated in Malek Mohammadi and Haji Zeinolabedini (2018) the most frequent keywords in Knowledge Organization theses were cataloging, information retrieval, library software, IRANMARC and indexing. On the other hand, one of the most significant events in Iran is the annual congress of Iranian Information experts. This event is holding every year in Iran since 2015. Also, a winter school for Knowledge Organization will be held at Shahid Beheshti University this year. The main topics of this winter school are conceptual models (such as LRM and CIDOC CRM), content and metadata standards (such as RDA and BIBFRAME), and Knowledge Organization tools (such as ontologies and semantic systems).

3.0 Conclusion
Knowledge Organization has a changing nature and has undergone an evolution in cataloging in the last years. As stated in Fattahi (2007) “publishing of many textbooks, manuals, and articles is strong evidence for the significance of Information Organization in teaching and practice in Iran”. We need skilled students who are ready to work in Knowledge organization departments of different agencies and also we need a more emphasize on new developments such as semantic web. Therefore, Iranian libraries have to replace their traditional catalogs with new approaches. Likewise, there are some developments in the library software of Iran. For instance, a joint working group between the Central library of Shahid Beheshti University and the Central library of Allameh Tabatabaei University has been made to apply RDA in their library systems. Likewise, in Qoqnus software, CIDOC CRM has been implemented. In this software all necessary classes and properties are utilized to represent cultural objects. This project represented...
as the outcome of Niknia (2019) in modelling archaeological literature and it is available at https://qoqnus.nosa.com/nhims/search.

References
Approach to Domain *Community Health* and its Implications for Information Management

**Abstract:**
The aim of the research was to explore the domain community health to discover its thematic and social dynamics, in order to build a Knowledge Organization System. Methodology: Three of the methods described by Hjørland were used: research in indexing and retrieving specialties, guides to specialized sources of information and bibliometrical studies. Results: community health is continuously used in specialized documents, so it has literary warrant. However, it is not included in KOS as a descriptor. This is an interdisciplinary emerging domain characterized by a low consensual terminology and an uncertain and scattered set of topics involved in, representing it.

1.0 Introduction
Community health is a specialty of recent emergency with interdisciplinary features. For this reason, little work has been done to know about its characteristics from the perspective of knowledge organization for information management. Thus, it was necessary to know its internal composition, conceptualizations and evolution, to understand the domain and its dynamics to create, as soon as possible, a Knowledge Organization System (KOS) for this specialty.

2.0 Aim
To reach a consensual concept of community health, to figure out the structural and social dynamics of the domain, to find out the set of terms that represents it and its thematic composition.

3.0 Methodology
The methodology chosen for the study was domain analysis, developed by Hjørland and Albrechtsen (1995). Hjørland (2002) proposed eleven methods of domain analysis. In this case, three of them were used:

a) Research in indexing and retrieving specialties. An exploration of the term community health in KOS and databases was conducted in order to find out the acceptance of the term.

b) Guides to specialized sources of information. A review of four databases was done, in order to find journals related to the domain, analyze them and identify the topics that represent the field of knowledge community health. This information set light to the interdisciplinary nature of community health.

c) Bibliometrical Studies. Sixty-one journals retrieved in the study described in b) were searched in the Web of Science. From this search 31 journals were found and 20115 articles from the period 2000-2019 were extracted and processed by the software tool for the analysis of the evolution of scientific knowledge, Sci-MAT.
4.0 Results
Through the application of the described methodology, the following results were obtained:

a) Research in indexing and retrieving specialties. After checking the acceptance of the term community health in the sources consulted (KOS and databases) the following results were obtained:
   - The term is continuously used in specialized documents and is recognized and used by specialists. The expression has literary warrant.
   - On the other hand, most of the KOS and databases consulted do not include it as a descriptor (indexing or search term), except for the Medline search system.

b) Guides to specialized sources. Sixty-one journals of community health could be retrieved. From them, the following results were found:
   - only 17 of the 61 journals showed integration of knowledge. The direct source for the identification of the topics and the interdisciplinarity of the specialty were the titles and scope of the journals, which means 28% of them. The rest of them present a disciplinary approach to community health, but neither integrate knowledge nor erase disciplinary boundaries.
   - A set of topics representing community health could be identified, which allowed the following classification proposal to be made: a) Scopes and phenomena of society; b) Human groups; c) Fields of Knowledge and Professional Practices; d) Health Programmes; e) Application places; f) Health problems.

c) Bibliometrical Studies. The results are the following:
   - The processing of these records led to the obtaining of a group of terms made up of the keywords given by authors and sources, which represent a first approximation to the terminology of this domain. Regarding this terminology, we found that around 43% remain along the time period studied. Therefore, it is possible to say that this set of terms constitutes the nucleus of the domain. It is also possible to appreciate a considerable dynamic in terms, which shows a young interdiscipline. This result shows what happens in young interdisciplines, which Szostak, Gnoli and López-Huertas (2016) call alluvium terminology.
   - A diachronic (evolution) map 2000-2019 could be made. It allowed identifying a set of subjects representing the domain that were classified into the multidimensional representation proposed in the preceding step.
   - The main subjects representing Community Health domain, were the following: odontology, psychiatry, public health, quality of life, nursing care, physical education, psychology, social service, cardiology, and primary health care.

5.0 Conclusions
The following conclusions have been reached:
   - The existence of a differentiated field of knowledge called community health was verified. On the other hand, KOS do not include it as a descriptor of search term except for Medline.
- The domain analysis methods used were adequate to obtain the intended results because they allowed to measure the domain of a relatively young field of knowledge.
- At present, community health is not a domain that integrates its disciplines; therefore, it cannot be said to be an interdiscipline in the strict sense. Currently, we are facing a domain in which multidisciplinarity predominates, which is a characteristic of young interdisciplines.
- It has been possible to know the thematic composition of the community health domain and the weight that each of them has within the field of knowledge.
- Nuclear and secondary terminology representing community health has been identified.

References
Provenance as an Ethical Measure for the Archival Knowledge Organization of Photographs

Abstract:
This paper reflects on the use of the principle of provenance in the archival knowledge organization of photographs as a critical an ethical measure to fight their misuses. It discusses the concept of provenance from a pragmatist and domain-analytic perspective to assess the potential of this principle in the determination of the contextual information of photographs to counteract fake news.

1.0 Introduction
In critical Archival Science, photographs that are produced for institutional purposes have achieved a fundamental role due to, among other things, their evidential value and potential to construct emancipatory narratives. In the present paper, we aim to explore the importance and possibilities of the principle of provenance as a pragmatist way of fighting injustices and fake news in the context of archival knowledge organization. More specifically, we aim to further recognize the role of the principle of provenance as a domain-analytic technique (Guimarães and Tognoli 2015) that can be used as a device for the study and development of knowledge organization systems and descriptions in which images taken out of context can be used to reinforce bias or unjust perspectives.

2.0 Domain analysis, provenance, and photographs in archives
According to Hjørland (2016), knowledge organization in archives should be considered part of the Knowledge Organization field while the principle of provenance is its most important principle of organization. Although there are multiple definitions and meanings of the principle of provenance (Tognoli and Guimarães 2019), we argue that the principle of provenance, while proposing the organization of groups of documents based on their production context, can serve as an ethical measure for the semantic representation of photographs. We believe that from Hjørland's post-Kuhnian perspective, the view of the domain established in the provenance could be compared to the scientific practice in the laboratory, as studied by Latour and Woolgar (1979). Here the production of facts in the laboratory would be similar to the production of knowledge in the archive of an institution. In this sense, documents in archives evidence the actions that originate them, while at the same time the principle of provenance dictates that documents must be organized in a way that reflect the activities, values, and interests of the institution. While the declassification of archives (Garcia Gutiérrez 2007) is a necessary step towards the realization of democracies and open societies, the application of the pragmatist principle of provenance and other critical techniques in the representation of images in knowledge organization systems, information systems, catalogs, and mass media ar-
chives (García Gutiérrez and Martínez-Ávila, 2014) can be used to dismantle the pre-
sumption of objectivity that underlies their political use in fake news in the post-truth
society.

The image that the photograph conveys should be considered an index of the produc-
tion process and not a unique element of organization. Thus, in order to make the dia-
logue between the context of production and its use in a way that is effective and con-
sistent with the functions of the archival document, we propose that the archival
knowledge organization processes and, in particular, the organization of the knowledge
evidenced by photographs must be ruled by the analysis of each domain according to
the principle of provenance and following the institutional process of declassification.
As a result, we propose that each institution should be analyzed as a specific domain in
which the discursive factors already established in the archival domain should also in-
clude the praxis of the institution (as the praxis in the archival domain corresponds to
the different processes of the division of labor in the archive, according to the legal and
administrative requirements, functions, and activities that become evidence in the gen-
eration of documents).

3.0 Conclusion

We believe that since the information recorded in the archival document refers to the
activities performed by the institution, it would be more appropriate to emphasize the
aspects related to the practical activities and not only the communicative aspects. The
reason for this is that the archival empirical practices and reflections emerge based on
the principle of provenance, looking for the identifications of the context of production
of the documents, interpreting their organic storage, and establishing their external and
internal organization as the basis for the functions of the archive (including the classifi-
cation and the description/representation). Thus, the principle of provenance is a vital
process for the photographs as it differentiates and highlights the domain of the archival
knowledge organization in relation to other forms of knowledge organization.

With the provenance and context of the institutional production that originated the
photograph, it would be possible to preserve, to some extent, the original intention of
the photographic record production (or at least to avoid the implicit misuse of its semi-
otic features) and avoid the manipulations linked to prejudices and social injustices.
Moreover, it can guarantee the diversity of knowledge domains, since each context of
document production presents its own specificity of intentions and social actions

References
Clasificación. Barcelona: Anthropos
García Gutiérrez, Antonio and Daniel Martínez-Ávila. 2014. “Critical Organization of Knowledge
Guimarães, José Augusto and Natália Bolfarini Tognoli. 2015. “Provenance as a Domain-Analysis
Tognoli, Natalia and José Augusto Guimarães. 2019. “Provenance as a Knowledge Organization
Phylomemetic Cataloguing
Expanding Bibliographic Relationships Beyond FRBR

Abstract:
This poster will outline the design of a new library cataloguing model, a design based on the relationships between texts. Previous efforts within knowledge organization have resulted in impressive systems, but the present moment presents an opportunity for extending the capabilities of catalogues to better incorporate user engagement. This can be done with a synthesis of Knowledge organization across many domains, incorporating concepts from fields as diverse as Biology, Philosophy and Literary Theory: Phylomemetic Cataloguing. The growing demands put upon catalogues by their users present cataloguers with an opportunity to expand the functionality of their catalogues. There has been an enduring interest in bibliographic relationships. (Niu 2013; Noruzi 2012) However, to date there have been few efforts to use phylogenetic relationships in the examination of works. Capturing the relationships between texts is conceived as a line of descent of ‘textual memes’. These relationships would capture user perceptions of the descent of memes and can be visualized as family trees of related texts.

1.0 Cataloguing Background
The current library model, Functional Requirements for Bibliographic Records (FRBR), was adopted when the new cataloguing guidelines, Resource Access Description (RDA) replaced the Anglo-American Cataloguing Rules Second Edition (ACCR2). While RDA includes rules for creating relationship links between different texts, such as sequels, references and adaptations, its FRBR-based structure encourages cataloguers to be more concerned with the connections between different versions of the same work, such as editions and translations. FRBR makes certain philosophical assumptions, namely that any particular text can be best described in terms of a Four-level entity-relationship model, connecting item to manifestation, manifestation to expression, and expression to work. Each of these groupings exists at a certain level of abstraction and can contain many of the lower groups. The work represents a sort of Platonic Ideal of the text, a pure form that becomes more and more particularized as we move down to expression and manifestation by the vagaries of a particular language, the hand of a particular author or the quirks of a particular printing press. These vertical connections prioritized, as a result, despite RDA’s inclusion of all sorts of possible relationship descriptors, the primacy of FRBR means that cataloguers do not capture all the important relationships between catalogued items.

2.0 Cladistics and Phylogenetics
Since the work of Carl Linnaeus in the 1750s biologists have made considerable progress in defining the often messy relationships between various organisms. Biologically inspired subject classification has a long tradition in LIS going back at least as far as the 1940s with the US Army Medical Library (Chan 1994). Bibliographic relationships are similarly organic, and that the insights of biologists can assist us to generate relationship trees based on a variety of factors, such as new editions, sequels, inspired works and academic citations. This would be of great use to readers, writers and researchers.
Biological taxonomy has a rich range of possible relationships that might be applied to information relationships. Cladistics, the dominant paradigm in taxonomy, is primarily concerned with lines of descent. However, with greater understanding of phenomena such as horizontal gene transfer and hybridization, modern taxonomy is able to describe many types of relationships that would serve as a model for the process of capturing and displaying the relationships between texts. This more granular and nuanced conception of descent is called Phylogenetics. Phylogenetic trees, as developed out of cladistics in biology, can be used as a model for bibliographic cataloguing systems. These trees have an added utility for importing to cataloguing, they lend themselves well to visualization.

3.0 Memetics

Genes are the units of inheritance in biology; memes are the units of inheritance in culture. Just as genes can be incredibly diverse in function and manifestation, memes can be similarly varied: simple phrases passed from parent to child, such as “an apple a day keeps the doctor away”, traditions, rituals, and best practices. Like genes, memes have a line of descent; however this project will only focus on ‘textual memes’, those units of cultural inheritance that manifest in ‘texts’. To be clear, texts in this case is to be understood in the broad sense, including more than just traditionally published books.

Another way to understand textual memes is through the example of literary tropes that manifest in the creation of texts. Such tropes include plot structure, character archetypes, and genre conceits. Since there are more texts than just narratives in many catalogs it is insufficient to focus solely on narratives; tropes are a case study of narrative ‘textual memes’. The development of and connections within music, art, science, and more could all be captured with a suitably robust cataloging system based on the mapping of textual memes beyond literature tropes.

4.0 Bibliographic Relationships

There has been an enduring interest in bibliographic relationships (Niu 2013; Noruzi 2012; Riva 2004; Smiraglia 2005; Tillet 2003). However, to date there have been few efforts to use phylogenetic relationships in the examination of works. According to Philipp A. Maas (2009) “a branching diagram that reflects the transmission history of a given text as truthfully as possible, is of fundamental importance for critical editing, since it enables the editor in many cases to judge the historical relationship of different text versions.” Despite Maas’s ability to analyze the Carakasamṛhitā Vimānasthāna, a classical Indian text with a long history of modification, there has not yet been a concerted effort to map the the standard relationship descriptors between works onto the phylogenetic model. This could be accomplished through a modification of the current cataloguing rules that focus on the relationships between works. From these relationships coherent groups of common descent, or clades, of related material would emerge. These clades would then go on to serve as improved localized finding aids.

This project resonates with emerging transformations of library cataloguing; the current FRBR standard has laid the groundwork for the new Library Reference Model (LRM) (IFLA 2017). The LRM envisions a linked data approach to library cataloguing, with a consequent shift of emphasis towards semantically-encoded relationships between texts at various levels of FRBR. Each of these developments is a step towards
being able to capture more and more of the cultural and social context in which creative works arise (Coyle 2016).

5.0 Phylomemetics

A functional prototype catalogue will use the Phylogenetic model and a select corpus of texts that contain ample and well documented inter-relationships at the monograph level, the level of particular editions. They will be catalogued first with the emerging linked data standards, such as the Web Ontology Language (OWL) based on phylogenetic approach, which will allow the visualization of these relationships in phylogenetic trees (Letunic and Bork 2011). This will serve to establish the viability of the phylogenetic approach to cataloguing, building on the emerging linked data approach as it built upon earlier traditions.

The immediate goal of this new cataloguing model is to establish the parameters of a new bibliographic model that could dramatically enhance the library community’s exploitation of linked data in its catalogues. The impact of this new cataloguing system would be enormous, allowing users to navigate organically from one text to another. While the ability to search for a text by author, topic or other category will remain, the additional relationship links and the possibility to create visual relationship maps will be of great use to many people.

The future of information appears to be increasingly focused on the connections between entities. If libraries and other information institutions wish to capitalize on this emerging paradigm they must create new tools. Phylogenetic-inspired cataloguing can be such a tool.

References
Abstract

D. J. Foskett was one of the first foundational authors to raise concerns about the cultural aspects of the Knowledge Organization. This paper investigates the contributions of this author to the Knowledge Organization cultural studies, aiming to identify through the literature the extent to which his contribution has influenced Knowledge Organization Systems regarding cultural aspects. In this sense, going into the scope of our study we are committed to discussing: What are D.J. Foskett's influences on the construction of modern knowledge organization systems? The theoretical and methodological foundation of our research lies in the historical, philosophical and epistemological foundations of the Knowledge Organization.

1.0 Introduction

The research agenda of the Knowledge Organization (KO) has privileged studies and research on issues involving socio-political, ethical, cultural and technological issues regarding the representation, storage and dissemination of the knowledge produced, in order to make patent the information practices that are, social practices par excellence, involving actors who play different roles in the cycle of apprehension, production and use of knowledge to meet and satisfy the most diverse information needs to account for their personal and professional activities in contemporary society. From this perspective, we seek to establish points of interlocution between the actors involved in the production of knowledge, the construction of knowledge organization systems (KOS) and their reflexes in the knowledge representation and the information retrieval in contemporary society from the studies of Douglas John Foskett (DJF), considered here as one of the foundational authors of cultural studies within the KO.

Cultural Studies, as an academic discipline emerged in the 1970s, which corresponds to the "cultural turnaround" of the Social and Human Sciences and in the same movement contributed to the destabilization of the frontiers of older disciplines such as History, Sociology, Literature, among others. Cultural Studies work as an agent in the reconfiguration of the disciplinary structure of the Humanities and Social Sciences, in a process that is still ongoing today (Baptista, 2009).

2.0 The method

Hjorland (2002) presents eleven methods for domain analysis. These are: production of literature guides or subject gateways; constructing special classifications and thesauri; indexing and retrieving in specialities; empirical users studies; bibliometrical studies; historical studies; document and gender studies; epistemological and critical studies; terminological studies, languages for special purpose, databases semantic and discourse studies; structures and institutions in scientific communication; scientific cognition, knowledge expert and artificial intelligence. Domain analysis provides a guide to identifying a domain and a discursive community.
The theoretical and methodological basis of our research lies in Domain Analysis (Hjørland, 2002) in search of the historical, philosophical and epistemological foundations of the KO that will be used as the framework for the configuration and time-space analysis of the KO before D.J. Foskett, during his activities and the influence of his contributions to KO. To verify the production of knowledge in KO in the Social Sciences and its representation in the SOC from the cultural studies of KO based on the contributions of D.J. Foskett we will use the method of cartography (Rimbert, 1968).

3.0 Results
The Douglas John Foskett’s contributions to the theory and practice of KO. A founding member of the Classification Research Group (CRG). Indeed, he was an active contributor to the CRG’s work. In 1955 the CRG had endorsed “the need for a faceted classification as the basis for all methods of information retrieval”. This principle was later endorsed at the Dorking Conference in 1957. It was in this context that DJF’s scholarship and research made its mark through his work in the study and dissemination of the theoretical concepts of Ranganathan. Putting facet analysis into practice, he developed a number of classification schemes. Among other important works are his Classification and integrative levels (1960), The Sayers Memorial Volume edited with Bernard Ira Palmer (1961), Classification and Indexing in the Social Sciences (1963) and Science, Humanism and Libraries (1964).

4.0 Final Consideration
From this perspective, we seek to establish points of interlocution between the actors involved in the production of knowledge, in the construction of KOS and its reflexes in the knowledge representation and in the information retrieval in con-temporary society from the studies of Douglas John Foskett, considered here as one of the foundational authors in cultural studies within the KO. In accordance with this study, Douglas John Foskett has contributed to the dissemination of the principle "the need for a faceted classification as the basis for all methods of information retrieval", the theory of integrative levels and the improvement of the KO in the field of cultural studies.

References
Baptista, Maria Manuel. 2009. “Estudos Culturais: O Quê e o Como da Investigação.” Carnets 1
http://journals.openedition.org/carnets/4382-
The Design Domain is Divided
Issues in Interdisciplinary Library Classification

Abstract:
This work explores how the two dominant library classification systems in the United States, Dewey Decimal Classification and Library of Congress Classifications portray the ontological positioning of design. Both classifications reveal design has no classification schedule of its own in either system, and is instead divided largely between Fine Art and Technology/Engineering as subclasses. This subsumption isolates design into ontological siloes, with significant implications for research and practice.

1.0 Introduction
An ever-expanding volume of knowledge organization research shows that classification systems express viewpoints. Bowker and Star (1999) show how systems express particular views about labor and identity, with significant implications from healthcare billing to racial segregation. Additional scholars have shown how library subject headings, classification, and controlled vocabularies express normative views about race, gender, sexuality, ethnicity, religion, and class (e.g., Olson 2002; Furner 2007; Adler, Huber, and Mix 2017; Howard and Knowlton 2018). As these knowledge organization system express perspectives, they establish an ontological worldview of the domain they describe. How do the two dominant library classification systems in the United States, Dewey Decimal Classification and Library of Congress Classification, portray the ontological positioning of design, and what does each classification reveal regarding perceptions of design as a domain?

1.1 Design as a domain
Design is a complex domain with competing views regarding its status. Simon (1996) considers design to be an applied scientific field. Others, such as Cross (2011), argue that design is a unique discipline, distinct from sciences or arts. Epistemological and ontological perspectives on design have changed dramatically in the past century: have library classification schemes kept pace?

2.0 The method
Hjørland (2002) lists examination of tools like classification schemes as one of eleven methods for domain analysis, suggesting that information professionals can better understand a conceptual space, such as a field of study or discipline, through examination of the knowledge organization systems used to describe and index documents in that space. This poster will report on a close reading and critical analysis of these two major library classification systems to reveal an overarching domain analysis of design.
3.0 Discussion of findings

Preliminary findings reveal that design has no classification schedule of its own in either system, and is instead divided largely between Fine Art and Technology/Engineering as subclasses. Moreover, the relegation of design to multiple levels of subclasses within various topics reifies the notion that design is a process or activity within other disciplines or domains, rather than its own discipline. Thus, design theories and practices--unique aspects that delineate design as a distinct epistemology--are systematically classified subordinate to the objects that design produces. Instead of supporting diverse repertoire-building within the broader design domain, this subsummation isolates design research into ontological siloes, minimizing exposure to additional alternate knowledge. Potential implications include a lack of interdisciplinary research and outside-the-box innovation, rendering design less visible as a domain of its own.

References
Analysis on Twitter of the Actors and Rumors around the Ebola Epidemic 2018-2019 in the Democratic Republic of Congo

Abstract:
The last major Ebola outbreak took place in 2014-2015. This epidemic affected mainly three West African countries (Guinea, Sierra Leone, Liberia) and killed 20,000 people. A number of articles have studied the rumors that circulated during this outbreak on twitter. (Jin et al. 2014; Fung et al. 2016). Ebola disease still rages today in the Democratic Republic of Congo since 1 August 2018. It made more than 2,050 deaths. It’s the second-worst epidemic in its history after West Africa. No studies have been conducted to determine who is communicating around this epidemic and what types of tweets / rumors are being broadcast? To answer this question, we conducted an analysis on Twitter via the Radarly® software, over a period from April 1 to July 07, 2019. The keyword Ebola and the #Ebola were used, with a filter on the french language. 17,282 tweets were collected and categorized via the software. The tool also extracted and represented knowledge in a cartographic way. It also helped to identify the dominant themes in the form of clusters. The tone of messages has also been determined. Our work has highlighted several actors communicating around the epidemic: general public, experts, politicians, media... Concerning the general public, it was highlighted that a part of the congolese population plays the role of relay of preventive messages and fights against fake-news. However, another party thinks that the disease is a plot to destabilize the country and spreads the rumor. Regarding the experts, they share many tweets to inform the general public. Concerning the politicians, the majority supports the medical teams and shows the example. In conclusion, our study, contrary to the 2014-2015 results, retrieve a congolese population divided in two: a part that admits the disease and adheres to the treatment and an another that considers it as a conspiracy. Our study suffers from bias. We did not take into account other media like Facebook and we limited our study to the French language.

1.0 Introduction

The last major ebola epidemic in Africa mainly affected three West African countries (Guinea, Sierra Leone, Liberia). It took place in 2014-2015 and killed about 20,000 people. A number of articles have investigated the rumors that circulated on Twitter during this epidemic. For example, Fung’s article pointed out that these media disseminated false information concerning the treatment of the disease, such as bathing in salt water to cure (Fung et al. 2016) Jin's article also pointed out that these media was behind fake news of a snake at the origin of the epidemic (Jin et al. 2014).

Ebola virus disease is still raging in the Democratic Republic of Congo (DRC) since August 1, 2018. It has killed more than 2,050 people. It is the second largest epidemic after the West African epidemic (Medley et al. 2020); Numerous studies have shown that Twitter is used by public health organizations, in particular to inform or educate the population, particularly in the event of a disaster or a sanitary crisis (Hart et al. 2017). However, no studies have been conducted to determine who communicates about the current Ebola epidemic in the Democratic Republic of Congo (DRC) and what types of tweets / rumors are circulated? To answer this problematic, we have conducted an analysis on Twitter over a period from 01/04/2019 to 07/07/2019. The keyword Ebola and #Ebola were used, with a filter on the French language. 17,282 tweets were collected and analyzed by a software (Radarly®). This tool permitted to
represent the knowledge in a cartographic way (volume of publications / time ...). It also permitted to identify the dominant themes in the form of clusters. The tone of the messages has also been determined. This article presents the main results found in our analysis, in particular the fact that several actors communicate around the epidemic.

2.0 Methodology

We used the Radarly® social media monitoring software marketed by Linkfluence (https://radarly.linkfluence.com/login) (Figure 1).

![Radarly® Interface](image1.png) ![Cartographic representations](image2.png)

This software allows to collect data on the social web (Tweet, Facebook, Instagram, forums, blogs, etc.). The software also makes it possible to represent the results in a cartographic manner, in particular in the form of clusters of dominant subjects. It also makes it possible to carry out analyzes of the tone of the messages published. It also identifies "influencers". To collect data from Radarly® software, we applied the monitoring process and the methodology developed by Tanti et al. (2012) and which includes 6 steps: definition of monitoring themes, identification, collection, analysis, synthesis and distribution of documents. We have analyzed tweets posted on Twitter only over a period from 04/01/2019 to 07/07/2019. A total of 17,282 tweets were collected, classified and categorized via the software. The software also enabled cartographic representations of the volume of publications as a function of time (Figure 2), making it possible to deduce media peaks intimately linked to health events.

3.0 Results

The requests have identified the main actors who communicate on Twitter concerning the Ebola disease which has been raging since 01/08/2018 in the Democratic Republic of Congo (DRC). The filter on the French language made it possible to select only tweets written in French. The analysis also made it possible to identify the messages conveyed. The main players found are: the general public, mainly Congolese citizens and associations; experts and health organizations (Ministry of Health of the Congo, WHO, etc.); the press, mainly Congolese; and politicians, mainly Congolese.

Concerning the general public: It is mainly congolese citizens and associations who speak about the epidemic. In the congolese population, there are divided opinions and two populations: a population that « believes » and a population that does not « believe » in the disease. The party that "believes" accepts the disease and the epidemic and considers it as a public health problem. It adheres to medical treatment and preventive measures. It relays messages of scientific information, education and awareness...
rest of the population is less "gullible" and denies the disease. It constitutes the majority of the tweets found (high negative tone). Thus, despite the efforts of response teams since the start of the epidemic in August 2018, this population considers the disease as a plot to destabilize the country. This population spreads the rumor. She accuses, for example, the laboratories or the WHO of having created the virus. Concerning the experts: It was mainly the national and international health organizations responsible for the response to the disease who spoke on Twitter during the study period. In particular, we observed that they shared many tweets to inform the general public. For example, the Ministry of Health of the DRC made a regular update on the disease which it relayed on Twitter. Concerning the media: It is mainly journalists and the press, both congolese and international, who speak. They tend to share WHO response releases, WHO prevention and awareness messages. Concerning the politicians: It is mainly congolese politicians who speak out. They relay in particular prevention, health education or awareness messages. For example, a tweet relaying the photo of the President of the Congolese Republic (H.E.M. Felix Tshisekedi) who complies with medical requirements during his national tours has been relayed several times on the media. In conclusion, our study thus found 12 main influencers and highlighted a negative message tone of 73.31%.

4.0 Conclusions

Our study, unlike the results found in the 2014-2015 epidemic, highlights a divided congolese population. Some people accept the disease and adhere to the treatment and another party views it as a conspiracy.

Our study suffers from limits. The short analysis period, like the choice to limit to the french language, is questionable. In addition, in our analysis, we did not take into account another media such as Facebook and forums. Finally, it is especially the choice of analysis on the social network Twitter itself which is questionable. Indeed, this media limits the number of characters present in messages. It thus limits long discussions, making it therefore the relay of current events and the engine of polemics and debates rather than the federator of true micro-communities.

References


Knowledge Organization Systems as Accountability Tools in Archival Science

Abstract:
Transparency and accountability are two important concepts in democratic societies which development parallels the evolution of knowledge organization. In archival science both can be found in classification schemes – understood as knowledge organization systems - that can be perceived as a mirror of the entity’s functions and structures. In this paper, we present an ongoing research project that aims to discuss knowledge organization systems as accountability and transparency tools in archival science.

1.0 Introduction
This poster presents an ongoing research project that aims to discuss knowledge organization systems as accountability and transparency tools in archival science. In this initial phase, we aim to confirm that a classification scheme, understood as an archival knowledge organization system, can offer a better comprehension than legal provisions (i.e., access or transparency laws) to accountability and transparency because it reveals an action that has occurred, while legal provisions cover situations that may occur.

2.0 Discussion
Records are a necessary and required product for the functioning of any organized entity, playing an important role in the transparency of actions, especially in democratic societies that are committed with administrative and historical accountability. Menne-Haritz (1994, 541) attests that “no law can be strong enough to make people do something that has no meaning for their activities. They create records because they need them, not because someone ordered their creation”.

Recently, there has been a fertile ground for accountability discussions on archival science. According to Dirks (2004, 32) “[…] for organizations, this accountability has meant a need to meet effectively and efficiently their mandates within the legal, cultural, and political climate in which they operate. For the public, this has meant a call for greater transparency, spawning the passage of freedom of information, protection of privacy, and other sunshine-oriented legislation”.

Accountability is an intersection point between archives and democracy. Miguel (2005) defined two axes of the concept: horizontal accountability, which means the control the established powers have over each other, and vertical accountability as the need of the representative to account and to be submitted to people’s verdict.

In this context, the archives play a central role in these mechanisms for controlling and ensuring accountability, representation, and responsiveness of representatives. If, from the principle of legality, we have that “the Public Administration can only do what the law allows ” (Di Pietro 2014, 65) the records remind us “that a law does not act; it is only a real human being who acts” (Hegel 1991, 178).

Transparency and accountability are, therefore, two important concepts in democratic societies which development parallels the evolution of knowledge organization (Smiraglia 2014). The latter holds in its core classification as a fundamental activity perceived as a process of distinguishing and distribution kinds of "things" into different
groups (Hjørland 2017), relying on structure to reveal the relationships that govern an ontological reality (Smiraglia 2014).

In the archival domain, a classification scheme – understood here as a “knowledge organization system designed for organizing knowledge and information, and making their management and retrieval easier” (Mazzochi 2018, 55), must be built according to the network of relationships between the record, its creator, the activity that created it, and the other records belonging to the same documentary set.

These relationships when represented in a classification scheme based on the entity’s function or structure elucidate what Duranti (2015) calls the archival bond (or what Smiraglia (2014) called ontological reality.

Therefore, a function-based classification scheme can go beyond the established relationships between records because it goes deeper, revealing the links between the record and the competences and activities of its producer. In other words, a classification by function is based on the context of a record’s creation and use, rather than on the content of the record itself (National Archives of Australia 2003).

The following is an example of a function-based classification scheme elaborated by The Public Archives of the State of Rio de Janeiro (APERJ). In the example, the classification scheme goes down to the record’s level, due to the use of a methodology called archival identification, which allows the archivist to get to the level where the activities are accomplished.

13 – Competence: To coordinate the Government planning and budgeting
13.01- Function: Human resource management policies
13.01.01 Activities: To analyze employees rights and benefits
13.01.01.01 Records typology: Statement of position accumulation
13.01.01.02 Records typology: Request for salary allowances

In the example above, the classification scheme is used as a transparency and accountability tool once it shows how the records are connected with their creator and how they are a product of the relationships between competences, function and activities.

Therefore, archival knowledge organization systems occupy a privileged position concerning accountability, since they shed light on the relationships between the record, the set it belongs to, and its producer.

3.0 Conclusion

Based on preliminary results, this poster aimed to discuss how archival knowledge organization systems, specifically classification schemes, could be conceived as a tool for transparency and accountability in democratic societies.

The main idea was to demonstrate that legal provisions could not be thought apart from knowledge organization systems once archivists have the methodological tools (Menne-Haritz 1994) to make evidence and information accessible, ensuring that the context is represented in a way it could be used to ensure accountability.

References


Autism Disorder in KO
Classification, Representation and Social Impact

Abstract:
In the health domain, the limitations and uses of language for the knowledge representation reveal a few dilemmas that extrapolate the linguistic structure: building different epistemological and ontological approaches materialized in knowledge organization systems. From the philosophy of language, this work analyzes the classifying condition of Autism Spectrum Disorder between the health domain and information retrieval, discussing the intersubjective dimension that crosses the meta-representation of knowledge-organization instruments. The complexity associated with this domain, which is present in the medical debate about the concept definition, refers to the spectrum of the possibilities to recognize the autistic individual. This spectrum guides us to the pragmatic dimension in the epistemological construction of the knowledge organization. It also finds its roots in the semiotic studies domain of the limits of the Theory of the Concept. Even though the Portuguese version for the Diagnostic and Statistical Manual of Mental Disorders does not use the terms "severe" or "mild" to classify the degrees of autism, there is evidence of them being recurrent on digital content found on the web. From an empirical perspective, this research aims to verify if and how the languages used on the web, specifically the use of mild and severe dichotomy, affect the construction of official discourses about the degrees of autism and their forms of representation. It is a quantitative and qualitative analysis segmented in three steps: analysis of the DSM e International Classification of Diseases (ICD), analysis of the scientific production about autism indexed in the Web of Science, Scopus, LILACS and Medline databases in the last five years (we will only consider documents indicating the use of the term mild as a degree of severity of autism) and evaluation the unscientific production published on Twitter from October to December 2019. The results aim to highlight the gaps in the representation of autism and identify promising theories of knowledge organization for understanding these dilemmas.

1.0 Introduction
The manifestation of oneself depends on a combination of physical and symbolic language. In a Cratylus dialogue, Plato debates the nature and arbitrariness of names and the possibility of language encompassing the exact representation of the object itself, that is, whether it would reflect the object itself. This doubt is intrinsic to the knowledge organization (KO). Its theoretical and epistemological demarcations are based on recorded information, which is already materialized into words and symbols.

If what determines the subject is what he thinks and how he expresses himself, how is it possible to name and classify individuals who may not express or interpret codes (such as the native language, for example) in the same way as their peers? This is the reality faced by Autism Spectrum Disorder (ASD), which impacts on several fields of knowledge. Without unique genetic markers, it has different effects, sometimes contradictory, depending on the individual. Based on the understanding that the human being is not constituted ontologically only by a biochemical combination, with contributions from the KO, anthropology, philosophy of language, and semiotics, the work investigates this question: is the classification "mild autism" possible? Specifically, this work explores the use of the word "mild," which affects the construction of official discourses on the degrees of autism and their forms of representation.
2.0 Methodological approaches

Hjørland (2008) explains that the KO is about not only documents, but also concepts and mediation of these concepts. This research sought to identify the political and social power of mediation through concepts in the context of autism.

The Diagnostic and Statistical Manual of Mental Disorders (DSM) outlines three levels of severity: "demanding support," "demanding substantial support," and "very substantial support" from two broad categories of analysis: social communication and restricted and repetitive behaviors. Stepping from DSM and International Classification of Diseases (ICD), quantitative and qualitative analysis segmented in the other two stages were performed at the practical level. The analysis of scientific production on autism indexed in the Web of Science, Scopus, LILACS, and Medline databases using the descriptors "mild autism" or "mildly autistic" in Portuguese and English totaling 764 documents. It was possible to observe an increase in scientific production with the use of terms from 2014. Thus, the choice was to reject the last five years (2015 to 2019). From the new total of 255, we selected those who debate degrees of severity and use the term mild autism, resulting in 31 documents. Despite the DSM and the ICD not providing for the classification of mild autism, in the scientific literature, researchers use this term, showing an increase in publishing in the last five years - which is interesting, given that the DSM-5 is from 2013. Therefore, its update does not seem to justify such popularization. The basis for the collection of terms published on Twitter, from October to December 2019, follows these search descriptors: (Asperger, 50, 199 items), ("autismo leve" OR "autista", 706 items), ("mild autism" OR "mildly autistic", 1,199 items), ("autismo severo" OR "autista severo", 910 items), ("severe autism" OR "severe autistic", 519 items), ("aspie" AND "autism mild", no results), ("aspie" and "mild autism", no results), ("Asperger" and "autismo leve", 44 items), ("Asperger" and "mild autism", 03 items), totaling 53,580 tweets. Since the goal was to verify how the use of certain terms can influence speeches, from a qualitative point of view, we decided to list the ten tweets of each descriptor with the highest number amount of shares and the 10 with the highest number of indication as favorites, totaling 123 tweets. Excluding the duplicates posts, it came to a total of 80 publications for conducting a content analysis. Of the 80 items, we identified 20 subjects, such as symptoms, prejudice, activism, request for help, and medicalization. For this work, the focus will be on two categories: against the use of the term mild autism (8 items) and use of the term mild autism (27 items). After analyzing the publications, we observed that of the 27 internet users who are in favor of "mild autism," 9 associate this degree with Asperger's Syndrome. It is interesting to note that, despite not having this information explicit in reference documents, such as the ICD and the DSM, the public tends to associate Asperger's syndrome as synonymous with a mild degree of autistic disorder.

3.0 Final considerations

The difficulties of classification in health multiply, therefore, within the limits of classification in the territory of the KO. The theoretical impact is in the difficulty that a classic conceptual approach contains in its formulation. A non-conceptual approach to the mediation between health classifications in the autism domain and KO
classifications for the same domain seems much closer to the tenuous political reality of autistic subjects.

All information production implies a recovery action for its sharing or elimination. In the social field, the KOS can define how people will live and which kind of medical services and public policies will be installed. In the beginning, we questioned if mild autism is a possible classification. When constructing the diagnostic categories, which questions are most relevant, which behaviors will be listed, which will be excluded, what weight will one behavior have to the other, which emotion will be seen as deviant, which will be the standard? The empirical research has shown that, gradually, the idea of mild autism is growing and building discourses on these individuals. This finding indicates the need for a more profound unfolding about the sources and user studies. The theories, methods, and instruments of KO should permanently dialogue and rethink their practices in the face of the autism domain and its challenges.

Acknowledgment
This study was funded by CNPq-Brazil.
Thanks to Josir Gomes (PPGCI IBICT UFRJ – Brazil) for helping us with the Twitter monitoring plataforma.

References
## Subject Index


Archival records, 352, 513

Archival representation, 339, 340


Aristotelian classification, 397, 520

Art, 30, 36, 39, 119, 145, 161, 162, 163, 169, 170, 198, 234, 263, 301, 309, 317, 318, 320, 381, 406, 430, 438, 439, 495, 498, 529, 564, 565

Arts, 109, 145, 146, 300, 301, 348, 352, 356, 388, 434, 479, 481, 483, 484, 488, 511

Authority control, 462

Authority data, 447

Authority record, 452

Authors, 42, 267, 460, 485

BARTOC, 52, 53, 54, 57, 58, 200, 201, 202, 203, 204, 205, 206

Basic Concepts Classification, 405, 421, 423, 427, 436, 438, 445, 527, 531

Basic Formal Ontology (BFO), 171, 173, 179, 502, 503, 504, 505, 506

Bias, 39, 98, 101, 103, 104, 236, 565

Categories, 126, 142, 146, 217, 225, 273

Classes, 18, 19, 257, 278, 455, 461, 521


Classifications, 52, 78, 150, 155, 200, 217, 219, 427, 430, 436, 461, 516

Cognitive work analysis, 333

Community health, 554

Community members, 210

Complex subjects, 66


Concept theory, 465, 466


Concepts and terms, 466

Conceptual map, 251

Conceptual models, 257

Content analysis, 35

Context, 77, 103, 104, 141, 160, 183, 198, 199, 374, 459, 473, 521, 539

Controlled vocabulary, 331, 514

Core elements, 551

Cultural aspects, 123

Cultural hospitality, 102

Cultural policy, 293

Cultural spaces, 324

Deep structure, 486, 487

Dewey Decimal Classification (DDC), 15, 35, 36, 39, 40, 50, 51, 78, 79, 80, 87, 142, 143, 144, 146, 147,
<table>
<thead>
<tr>
<th>Term</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>148, 149, 150, 221, 230, 234, 242,</td>
<td>331, 332, 421, 422, 427, 452, 455, 456, 457, 458, 461, 462, 466, 494, 496, 503, 520, 551, 552, 564</td>
</tr>
<tr>
<td>Dickinson Classification, 230, 234,</td>
<td>421</td>
</tr>
<tr>
<td>Digital libraries, 123, 408, 552</td>
<td></td>
</tr>
<tr>
<td>Domain analysis, 36, 557, 562</td>
<td></td>
</tr>
<tr>
<td>Domain trees, 322, 327</td>
<td></td>
</tr>
<tr>
<td>Encyclopædia Britannica, 142, 149,</td>
<td>150</td>
</tr>
<tr>
<td>Enterprise search, 328</td>
<td></td>
</tr>
<tr>
<td>Equivalent terms, 158</td>
<td></td>
</tr>
<tr>
<td>Facet analysis, 117, 256, 257, 405</td>
<td></td>
</tr>
<tr>
<td>Faceted classification, 115, 117, 254,</td>
<td>259, 262, 331</td>
</tr>
<tr>
<td>Facets, 160, 256, 258, 260, 262, 383,</td>
<td>410, 427, 478, 491, 493</td>
</tr>
<tr>
<td>Fiction, 190, 193, 198, 235, 337</td>
<td></td>
</tr>
<tr>
<td>FOAF, 236, 237, 238, 239, 240, 241,</td>
<td>242, 243, 438</td>
</tr>
<tr>
<td>Encyclopædia Britannica, 142, 149,</td>
<td>150</td>
</tr>
<tr>
<td>Enterprise search, 328</td>
<td></td>
</tr>
<tr>
<td>Equivalent terms, 158</td>
<td></td>
</tr>
<tr>
<td>Facet analysis, 117, 256, 257, 405</td>
<td></td>
</tr>
<tr>
<td>Faceted classification, 115, 117, 254,</td>
<td>259, 262, 331</td>
</tr>
<tr>
<td>Facets, 160, 256, 258, 260, 262, 383,</td>
<td>410, 427, 478, 491, 493</td>
</tr>
<tr>
<td>Fiction, 190, 193, 198, 235, 337</td>
<td></td>
</tr>
<tr>
<td>FOAF, 236, 237, 238, 239, 240, 241,</td>
<td>242, 243, 438</td>
</tr>
<tr>
<td>Gaius database, 375, 376, 377, 378,</td>
<td>379</td>
</tr>
<tr>
<td>Genre, 170, 190, 191, 195, 198, 199,</td>
<td>227, 230, 234, 235, 383, 424, 427</td>
</tr>
<tr>
<td>Graffiti art, 168</td>
<td></td>
</tr>
<tr>
<td>Horror films, 424</td>
<td></td>
</tr>
<tr>
<td>Human indexing, 375</td>
<td></td>
</tr>
<tr>
<td>Index, 22, 105, 110, 112, 113, 150,</td>
<td>234, 357, 358, 362, 385, 386, 387, 393, 394, 395, 520, 535</td>
</tr>
<tr>
<td>Index terms, 358, 362</td>
<td></td>
</tr>
<tr>
<td>Indigenous communities, 123, 124, 125</td>
<td></td>
</tr>
<tr>
<td>Indigenous peoples, 124, 128, 129, 131</td>
<td></td>
</tr>
<tr>
<td>Individual performance, 352</td>
<td></td>
</tr>
<tr>
<td>Information architecture, 333, 552</td>
<td></td>
</tr>
<tr>
<td>Information Coding Classification,</td>
<td>460, 462, 464, 478</td>
</tr>
<tr>
<td>Information retrieval, 147, 327</td>
<td></td>
</tr>
<tr>
<td>Integrative Levels Classification (ILC),</td>
<td>49, 52, 54, 56, 58, 423, 489, 490, 491, 493, 503, 506</td>
</tr>
<tr>
<td>Interface, 21, 123, 132, 226, 428, 433,</td>
<td>522, 527, 567</td>
</tr>
<tr>
<td>Interoperability with other vocabularies, 22, 30, 327</td>
<td></td>
</tr>
</tbody>
</table>
Opera, 42, 226, 232, 234, 235, 300, 481

Participant role, 486, 487
Performance information, 354
Performance materials, 350, 352
Personal expression, 212
Perspective, 41, 48, 78, 103, 189, 198, 215, 252, 272, 282, 327, 420, 446, 454, 469, 493, 501, 516, 565
Phylogenetic trees, 560
Polythetic groups, 217
Prison fiction, 195
Processes, 60, 415, 416, 417, 492
Provenance, 308, 446, 449, 453, 454, 557, 558

RDF records, 441, 443
RDF schema, 437, 441, 443
Relationships, 41, 42, 43, 44, 48, 69, 70, 72, 73, 75, 132, 140, 176, 223, 225, 252, 327, 445, 453, 468, 498, 559, 560, 561
Associative relationships, 229
Bibliographic relationships, 41, 42, 559
Hierarchical relationships, 229
Semantic relationships, 66

Schemes, 39, 87, 121, 122, 234, 336, 426
Semantic relations, 66, 245
Semantic technologies, 47
Semantic tools, 471
Semantic web, 156, 552
Signature pedagogies, 484
SKOS, 15, 20, 21, 49, 50, 51, 52, 57, 58, 158, 200, 201, 204, 206, 254, 255, 260, 261, 262, 263, 331, 336, 438, 439, 440, 443, 473
SKOS data, 20, 52

SKOS format, 15, 50, 57, 200, 440, 443
SKOS vocabularies, 200, 201
Social media, 404, 408
Social networks, 408
Social problem fiction, 195
Social sciences, 146
Social tagging, 191, 207
Subject analysis, 451
Subject headings, 155

Tags, 167, 193, 198, 199, 212, 213, 214, 215, 291, 511
Taxonomy, 48, 91, 92, 96, 180, 198, 225, 283, 300, 301, 331, 332, 333, 336, 337, 371, 373, 393, 561
Terminology, 16, 21, 22, 104, 201, 363, 434, 453, 461, 506, 518, 519, 521
Thesauri and interoperability, 22, 538
Time, 60, 71, 117, 215, 283, 292, 314, 317, 374, 393, 482
Tools, 17, 57, 60, 77, 170, 273, 309, 337, 393, 444, 459, 477, 521, 569
Twitter, 208, 566, 567, 568, 572, 573, 574

© Ergon - ein Verlag in der Nomos Verlagsgesellschaft

Understanding, 14, 16, 122, 171, 409, 435, 488

Universal Decimal Classification (UDC), 35, 58, 69, 78, 79, 80, 81, 87, 154, 155, 230, 234, 410, 426, 427, 436, 437, 439, 440, 441, 442, 443, 444, 445, 459, 461, 462, 466, 503, 520

User queries, 423, 530

User studies, 425


Version control, 289

Vocabulary mapping, 430

Warrant, 31, 33, 38, 39, 103, 131, 140, 198, 214, 328, 329, 334, 335, 336, 492

Cultural warrant, 102, 329

Literary warrant, 102, 285, 329

User warrant, 192, 329

Western art music, 226, 227, 232, 233

Words, 122, 131, 401
<table>
<thead>
<tr>
<th>Author</th>
<th>Page Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abgaz, Yalemisew</td>
<td>470</td>
</tr>
<tr>
<td>Adämmer, Philipp</td>
<td>374</td>
</tr>
<tr>
<td>Agustín-Lacruz, Maria Del</td>
<td>532</td>
</tr>
<tr>
<td>Andrade, Julietti de</td>
<td>537</td>
</tr>
<tr>
<td>Aracri, Giovanna</td>
<td>13</td>
</tr>
<tr>
<td>Araújo, Andre Vieira de Freitas</td>
<td>105</td>
</tr>
<tr>
<td>Araújo, Webert Júnior</td>
<td>23</td>
</tr>
<tr>
<td>Balakrishnan, Uma</td>
<td>455</td>
</tr>
<tr>
<td>Barber, Olatunde</td>
<td>507</td>
</tr>
<tr>
<td>Barité, Mario</td>
<td>31, 460</td>
</tr>
<tr>
<td>Barros, Thiago Henrique Bragato</td>
<td>236, 465, 549</td>
</tr>
<tr>
<td>Bawden, David</td>
<td>226</td>
</tr>
<tr>
<td>Bernaoui, Radia</td>
<td>540</td>
</tr>
<tr>
<td>Biagetti, Maria Teresa</td>
<td>41</td>
</tr>
<tr>
<td>Binding, Ceri</td>
<td>49</td>
</tr>
<tr>
<td>Blickhan, Samantha</td>
<td>133</td>
</tr>
<tr>
<td>Bochi, Fernanda</td>
<td>264</td>
</tr>
<tr>
<td>Buizza, Pino</td>
<td>59</td>
</tr>
<tr>
<td>Calatayud, Francisco José Morales</td>
<td>554</td>
</tr>
<tr>
<td>Campbell, D. Grant</td>
<td>69</td>
</tr>
<tr>
<td>Campos, Maria Luiza de Almeida</td>
<td>562</td>
</tr>
<tr>
<td>Caruso, Assunta</td>
<td>13</td>
</tr>
<tr>
<td>Carvalho, Priscila Ramos</td>
<td>522</td>
</tr>
<tr>
<td>Cheng, Yi-Yun</td>
<td>88</td>
</tr>
<tr>
<td>Clarke, Rachel Ivy</td>
<td>564</td>
</tr>
<tr>
<td>Colette, Djadeu Nguemedymam</td>
<td>543</td>
</tr>
<tr>
<td>Colombo, Stephanie</td>
<td>98</td>
</tr>
<tr>
<td>Correa, Lucas</td>
<td>569</td>
</tr>
<tr>
<td>Corujo, Luís</td>
<td>411</td>
</tr>
<tr>
<td>Crippa, Giulia</td>
<td>105, 394</td>
</tr>
<tr>
<td>Dias, Célia da Consolação</td>
<td>254</td>
</tr>
<tr>
<td>Díaz, José Luis Preza</td>
<td>470</td>
</tr>
<tr>
<td>Dorn, Amelie</td>
<td>115, 470</td>
</tr>
<tr>
<td>Duarte, Rodrigo Aldeia</td>
<td>546</td>
</tr>
<tr>
<td>Evangelista, Isadora Victorino</td>
<td>549</td>
</tr>
<tr>
<td>Farnel, Sharon</td>
<td>123</td>
</tr>
<tr>
<td>Folino, Antonietta</td>
<td>13</td>
</tr>
<tr>
<td>Frasisse, Amel</td>
<td>133</td>
</tr>
<tr>
<td>Fujita, Mariângela Spotti Lopes</td>
<td>532</td>
</tr>
<tr>
<td>Furner, Jonathan</td>
<td>142, 217</td>
</tr>
<tr>
<td>Galindo, Fernando</td>
<td>151</td>
</tr>
<tr>
<td>García-Marco, Francisco-Javier</td>
<td>151, 474</td>
</tr>
<tr>
<td>Gnoli, Claudio</td>
<td>49, 502</td>
</tr>
<tr>
<td>Gomes, Daniel Libonati</td>
<td>236</td>
</tr>
<tr>
<td>Gomes, Josir Cardoso</td>
<td>78</td>
</tr>
<tr>
<td>Graf, Ann M.</td>
<td>161</td>
</tr>
<tr>
<td>Guadarrama-Sánchez, Hugo Alberto</td>
<td>319</td>
</tr>
<tr>
<td>Guimarães, José Augusto Chaves</td>
<td>446</td>
</tr>
<tr>
<td>Hadi, Widad Mustafa El</td>
<td>181</td>
</tr>
<tr>
<td>Hashtroudi, Negin Shokrzadeh</td>
<td>479, 551</td>
</tr>
<tr>
<td>Haynes, David</td>
<td>171</td>
</tr>
<tr>
<td>Helfer, Olivia</td>
<td>455</td>
</tr>
<tr>
<td>Henry, Antoine</td>
<td>181</td>
</tr>
<tr>
<td>Hernández-Villeda, Rosa de</td>
<td></td>
</tr>
<tr>
<td>Guadalupe</td>
<td>319</td>
</tr>
<tr>
<td>Hider, Philip</td>
<td>190</td>
</tr>
<tr>
<td>Hindermann, Maximilian</td>
<td>200</td>
</tr>
<tr>
<td>Hoang, Khanh Linh</td>
<td>88</td>
</tr>
<tr>
<td>Holstrom, Chris</td>
<td>207</td>
</tr>
<tr>
<td>Ibekwe, Fidelia</td>
<td>264</td>
</tr>
<tr>
<td>Jung, Dasom</td>
<td>348</td>
</tr>
<tr>
<td>Kan, Rebecca Y.P.</td>
<td>484</td>
</tr>
<tr>
<td>Kelly, Elizabeth Joan</td>
<td>402</td>
</tr>
<tr>
<td>Kenfield, Ayla Stein</td>
<td>402</td>
</tr>
<tr>
<td>Khoo, Christopher S.G.</td>
<td>484</td>
</tr>
<tr>
<td>Kim, Seungchon</td>
<td>348</td>
</tr>
<tr>
<td>Kim, Yoonwhan</td>
<td>348, 357</td>
</tr>
<tr>
<td>Kleineberg, Michael</td>
<td>489</td>
</tr>
<tr>
<td>Koch, Gerda</td>
<td>470</td>
</tr>
<tr>
<td>Lara, Marilda Lopes Ginez de</td>
<td>537</td>
</tr>
<tr>
<td>Lasala, Pilar</td>
<td>151</td>
</tr>
<tr>
<td>Ledl, Andreas</td>
<td>200</td>
</tr>
<tr>
<td>Leazer, Gregory H.</td>
<td>217</td>
</tr>
<tr>
<td>Lee, Deborah</td>
<td>226</td>
</tr>
<tr>
<td>Lee, Hosin</td>
<td>348</td>
</tr>
<tr>
<td>Lee, Hyewon</td>
<td>348</td>
</tr>
<tr>
<td>Lee, Wan-Chen</td>
<td>494</td>
</tr>
<tr>
<td>Lima, Gercina Ângela de</td>
<td>23, 245</td>
</tr>
<tr>
<td>López del Ramo, Joaquín</td>
<td>151</td>
</tr>
<tr>
<td>López-Huertas Pérez, María José</td>
<td>554</td>
</tr>
</tbody>
</table>
PROCEEDINGS

Knowledge organization at the interface

16th International Conference
of the International Society for Knowledge Organization

International scientific Committee

David Bawden, City University of London, United Kingdom
Ann Bygholm, Aalborg University, Denmark
Amos David, Université de Lorraine, France
Stella Dextre Clarke, United Kingdom
Thomas Dousa, University of Chicago, United States
Jonathan Furner, UCLA-GSE&IS, United States
Claudio Gnoli, University of Pavia, Italy
Koraljka Golub, Linnaeus University, Sweden
José Augusto Chaves Guimarães, São Paulo State University, Brazil
Lala Hajibayova, Kent State University, United States
David Haynes, Edinburgh Napier University, United Kingdom
Birger Hjørland, University of Copenhagen, Denmark
Michael Kleineberg, Humboldt-Universität zu Berlin, Germany
Deborah Lee, City University of London, United Kingdom
Ying-Hsang Liu, University of Southern Denmark, Denmark
Maria López-Huertas, University of Granada, Spain
Haakon Lund, University of Copenhagen, Denmark
Marianne Lykke, Aalborg University, Denmark
Andrew MacFarlane, City University of London, United Kingdom
Devika Madelli, DRTC, Indian Statistical Institute, India
Jens-Erik Mai, University of Copenhagen, Denmark
Daniel Martínez-Ávila, Universidad Carlos III de Madrid, Spain
Widad Mustafa El Hadi, Université de Lille, France
Susanne Ørnager, University of Copenhagen, Denmark
Cristina Pattuelli, Pratt Institute, United States
Vivien Petras, Humboldt-Universität zu Berlin, Germany
Jian Qin, Syracuse University, United States
Pauline Rafferty, Aberystwyth University, United Kingdom
Athena Salaba, Kent State University, United States
Ali Shiri, University of Alberta, Canada
Mette Skov, Aalborg University, Denmark
Richard Smiraglia, Institute for Knowledge Organization and Structure, United States
Linda Smith, University of Illinois at Urbana-Champaign, United States

© Ergon - ein Verlag in der Nomos Verlagsgesellschaft
Dagobert Soergel, University at Buffalo, United States
Renato Rocha Souza, Fundação Getulio Vargas, Brazil
Raghavan Sundaratatachar, Sarada Ranganathan Endowment for Library Science, India
Tanja Svarre, Aalborg University, Denmark
Rick Szostak, University of Alberta, Canada
Joseph Tennis, University of Washington, United States
Martin Thellefsen, University of Copenhagen, Denmark
Marcia Zeng, Kent State University, United States
Maja Žumer, University of Ljubljana, Slovenia
Marit Kristine Ådland, Oslo Metropolitan University, Norway