Categories, Contexts and Relations
in Knowledge Organization
# Table of Contents

## INTRODUCTION

### KEYNOTE ADDRESS

1

## DOMAIN OF KNOWLEDGE ORGANIZATION

*Birger Hjorland.* Is Knowledge Organization = Information Organization?  
8

*H. Peter Ohly.* Mission, Programs and Challenges of Knowledge Organization  
15

*Rick Szostak.* The Basic Concepts Classification  
24

*José Augusto Chaves Guimarães, Ely Tannuri de Oliveira* and  
*Maria Cláudia Cabrini Gracio.* Theoretical Referents in Knowledge Organization:  
A Domain Analysis of Knowledge Organization Journal  
31

*José Augusto Chaves Guimarães* and *Joseph T. Tennis.* Constant Pioneers: The Citation Frontiers of Indexing Theory in the ISKO International Proceedings  
39

*Aline Elis Arboit,* *Maria Cláudia Cabrini Gracio,* *Ely Francina Tannuri de Oliveira* and  
*Leilah Santiago Bufrem.* Relationship Between Authors and Main Subject Categories in the Knowledge Organization Domain: A Bibliometric Approach  
44

## GENERAL CLASSIFICATION SCHEMES

*Almila Akdag Salah,* *Cheng Gao,* *Krzysztof Suchecki,* *Andrea Scharnhorst,* and *Richard P. Smiraglia.* The Evolution of Classification Systems: Ontogeny of the UDC  
51

*Joseph T. Tennis.* Facets and Fugit Tempus: Considering Time’s Effect on Faceted Classification Schemes  
58

*B.A. Sharada.* Ranganathan’s Colon Classification: Kannada-English Version ‘dwibindu vargiikaraNa’  
63

## KNOWLEDGE ORGANIZATION FOR THE DIGITAL ENVIRONMENT

*Carolyn Watters* and *Naureen Nizam.* Knowledge Organization on the Web: The Emergent Role of Social Classification  
69

*M. Cristina Pattuelli* and *Sara Rubinow.* Charting DBpedia: Towards a Cartography of a Major Linked Dataset  
75

*Christopher S.G. Khoo, Dong Zhang, Mi Wang* and *Xin Jie Yun.* Subject Organization in Three Types of Information Resources: An Exploratory Study  
83

*Kavi Mahesh* and *Pallavi Karanth.* A Novel Knowledge Organization Scheme for the Web: Superlinks with Semantic Roles  
90

96

*Evelyn Orrico,* *Vera Dodebei* and *Miriam Gontijo.* The Precision of Metaphor for Information Retrieval  
103

## KNOWLEDGE ORGANIZATION AS A NAVIGATION TOOL

*Charles-Antoine Julien,* *Pierre Tirilly,* *John. E. Leide* and *Catherine Guastavino.* Using the LCSH Hierarchy to Browse a Collection  
109
Kathryn La Barre and Rosa Inês de Novais Cordeiro. Unmasking „That Obscure Object of Desire“: A Brief Report from the Films and Facets Project


João Batista Ernesto de Moraes. Aboutness in Fiction: Methodological Perspectives for Knowledge Organization

Shu-jiun Chen, Marcia Lei Zeng and Hsueh-hua Chen. Alignment of Conceptual Structures in Controlled Vocabularies in the Domain of Chinese Art – A Discussion of Issues and Patterns

Miriam Gontijo, Vera Dodebei and Evelyn Orrico. Discourse Analysis as an Approach to Categorizing the Domain of Public Policy: The Case of Brazilian E-Government

KNOWLEDGE ORGANIZATION FOR ARCHIVES

Renato Rocha Souza, Flávio Codeço Coelho and Suemi Higuchi. The CPDOC Semantic Portal: Applying Semantic and Knowledge Organization Systems to the Brazilian Contemporary History Domain

Natália Bolfarini Tognoli and José Augusto Chaves Guimarães. Challenges of Knowledge Representation in Contemporary Archival Science

Thiago Henrique Bragato Barros and João Batista Ernesto de Moraes. Archival Classification and Knowledge Organization: Theoretical Possibilities for the Archival Field

Pekka Henttonen. Diversity of Knowledge Organization in Records and Archives Management

DESIGN AND DEVELOPMENT OF KNOWLEDGE ORGANIZATION TOOLS

Leonard Will. The ISO 25964 Data Model for the Structure of an Information Retrieval Thesaurus

Wiesław Babik. A Faceted Classification of Cartographic Materials: Problems of Construction and Use

Ming-Shu, Yuan, Fan-Hua, Nan and Gou-Chi, Lee. Constructing Knowledge Classification Scheme in Industrial Technology via Domain Analysis: An Empirical Study

B.L. Vinod Kumar and Khaiser Nikam. Sanskrit-English Bilingual Thesaurus for Yogic Sciences: A Case Study of Problems and Issues with Terms of Non-Latin Origin

Emilena Josemary Lorenzon, Luciana de Souza Gracioso, Marco Donizete Paulino da Silva, Marcele Tinelli, Roniberto Morato Amaral, Leandro Innocentini Lopes de Faria and Wanda Aparecida Machado Hoffmann. Controlled Vocabulary for Intelligence Information System for Shoes

INFORMATION MINING / AUTOMATIC INDEXING

Flávio Codeço Coelho, Renato Rocha Souza, Daniel Magalhães Chada and Pablo de Camargo Cerdeira. Information Mining and Visualization of Data from the Brazilian Supreme Court (STF): A Case Study
Carlos Alberto Corrêa and Nair Yumiko Kobashi. Automatic Indexing and Information Visualization: A Study Based on Paraconsistent Logics 326

Nalini A. Raja. Digitized Contents and Index Pages as Alternative Subject Access Fields 333

USERS AND CONTEXT

Carol L. Tilley and Kathryn A. La Barre. What if they build it and no one comes? Balancing Full-Text Access and User Tasks 338

Sholeh Arastoopoor and Rahmatollah Fattahi. Users’ perception of Aboutness and Ofness in Images: An Approach to Subject Indexing Based on Ervin Panofsky’s Theory and Users” View 345

Melodie J. Fox. Communities of Practice, Gender and Social Tagging 352

Radia Bernaoui and Mohmed Hassoun. User Expectations and Reality and Delineation of Agricultural Information Systems in the Maghreb 359

ABSTRACTS OF POSTERS

Abhijit Dasgupta and V K Thomas. Semantic Web Technology: A Key for Effective Knowledge Management: A Case Study on Facets of Law 365

Vera Dodebei, Miriam Gontijo and Evelyn Orrico. Knowledge Representation by Social identity Categories: The Brazilian Institute for Web Science Research, A Case Study 366

Daniel Martínez Ávila. Knowledge Organization and Application of Dublin Core Possibilities in CATIA V5 368

Jolanta Szulc. Building the Ontology Dictionary using the Inference Engine RACER 369


Michelli Sasaki, Michely Jabala Mamede Vogel and Nair Yumiko Kobashi. Aspects of Information Organization and Retrieval from a News Portal 374


Rahmatollah Fattahi, Mehri Parirokh, Mohammad Hosien Dayyani and Abdolrasoul Khosravi. The Efficacy of Google's Suggested Keywords and Phrases in Query Expansion Based on “The Least effort Principle" and "Cognitive Load Theory" 378


LIST OF CONTRIBUTORS 381

AUTHOR INDEX 389

SUBJECT INDEX 391

INTERNATIONAL SCIENTIFIC COMMITTEE 395
PROCEEDINGS

Categories, Contexts and Relations in Knowledge Organization

12th International Conference
International Society for Knowledge Organization

Introduction

The last time an International ISKO Conference was held in India was in 1992, in Chennai (then Madras). That conference - the 2nd International ISKO Conference - marked the culmination of a series of national and international events celebrating the birth centenary of Dr. S.R. Ranganathan. The conference theme and title of the volume was ‘Cognitive Paradigms in Knowledge Organization’. The present Conference – the 12th in the series – being held in India after 20 years marks Ranganathan’s 120th birth anniversary, a good coincidence! During the intervening two decades developments have been taking place that are impacting the knowledge organization domain:

1. Several genres and vast quantities of information and knowledge sources (e.g. people) are accessible at any time from anywhere via the Internet and many other international, national and local networks;
2. An increasing proportion of such sources are in languages other than those derived from Latin;
3. A growing number of information seekers and users wish to contribute to, access, browse and read information materials in their respective native languages or languages they are familiar with;
4. There is a growing recognition that different cultures and associated languages impact the perception, perspective, and frames of information / knowledge generators and seekers; and
5. Additionally, thanks to developments in international travel and globalization in almost all sectors of human activities there is increasing interaction and collaboration among people at different levels.

Information science literature of recent times – conference papers, journal articles and books - provide evidence that information professionals and specialists in cognate domains have been focusing on the above issues. Studies and thinking in this regard range from questioning the relevance, utility and value of such devices as subject classification schemes and thesauri in the digital environment to studies that have been specifically examining how such knowledge organization tools could be employed to enhance knowledge discovery and information retrieval in the digital environment. Also, innovations in and with the age-old devices are being created. The papers contributed to this conference which has as its theme, Categories, Relations and Contexts in Knowledge Organization too reflect these trends. Fifty five papers and ten posters are to be presented during the conference. The papers in this volume span several different but inter-related themes: Domain of Knowledge Organization, General Classification Schemes, Knowledge Organization for the Digital Environment, Knowledge Organization as Navigation Tool, Ontology, Categories in Knowledge Organization, Knowledge Organization for Specific Domains, Knowledge Organization for Archives, Design and Development of Knowledge
Organization Tools, Information Mining/ Automatic Indexing and Users and Context. The Internet and the World Wide Web have led to a situation in which many disparate communities, including the Knowledge Organization community within the Library and Information Science community, have seriously started addressing issues related to organizing knowledge to support information retrieval and knowledge discovery. It will not be incorrect to say that enhancements and improvements in Knowledge Organization will come about in the future mainly through a cross-fertilization of ideas, tools and techniques.

We would like to convey our sincere thanks to all those who helped make this volume possible. First and foremost our thanks are due to the authors who have contributed papers reporting results of their research; our thanks are also due to the members of the International Scientific Committee who reviewed papers and provided valuable inputs in the form of suggestions to improve the quality of the submissions to the conference.

Two special lectures organized by the Sarada Ranganathan Endowment for Library Science are being delivered as part of the proceedings of this conference; The Sarada Ranganathan Endowment Lectures 2012 and the CurzonCo Seshachalam Lecture. The text of the CurzonCo Seshachalam Lecture by Prof. Carolyn Watters appears in this volume as a paper. We hope to bring out the text of the Sarada Ranganathan Endowment lectures 2012 on ‘Sharing Information Resources across Communities and over Time: Metadata, Digital Libraries & Archives, and Information Schools’ by Prof. Shigeo Sugimoto as a separate volume.

We hope the discussions will be both interesting and productive and that you will also enjoy the cultural programme and visits to places of tourist interest.

Bangalore
22\textsuperscript{nd} June 2012

A. Neelameghan
K. S. Raghavan
Universes, Dimensions, Domains, Intensions and Extensions: Knowledge Organization for the 21st Century

Abstract
In Knowledge Organization (KO) there is work to solidify concept theory, which is at the core of our discipline; but there are other dimensions, as well as suggestions that classification must engage a multi-verse. This paper encompasses a domain analysis of KO as a means of visualizing the emergence and coherence of our domain, and as a way of denominating the parameters of the universe (or universes) in which our domain operates, as well as the dimensions of the operational paradigms at work. In other words, we look here at the extension and intension of KO as a domain. KO as a domain demonstrates coherence across time and across geopolitical boundaries, particularly as it concerns its theoretical foundations. Consistently marked dimensions within the domain: theoretical versus applied on one continuum, humanistic versus scientific on another. These dimensions serve to maintain constructive and dynamic tension within the domain, which in turn keeps the research front constantly in a state of renewal.

Introduction: Universes, Dimensions, Domains
In Knowledge Organization the domain KO, there is frequent mention of dimensions, particularly with regard to facets, and faceted classifications. Most such conversation is rooted in concept theory, and in most cases the dimensions put forward are merely categories (place, ethnicity, etc.), sometimes mutually exclusive, but not always. Yet it is not uncommon to perceive of the collectivity of knowledge (recorded or not) as a metaphor of the physical universe. And in the universe, there are many dimensions. Time and space are one, for instance, although most humans experience those as distinct.

In our domain at present there is work to solidify concept theory, which is at the core of our discipline; an example is Hjørland (2009). But there are other dimensions—yardsticks, as it were—in our work. And there are suggestions that classification must take into account not a single universe of knowledge, but rather a multi-verse (Heuvel and Smiraglia, 2010; Smiraglia, Heuvel and Dousa, 2011). Within these suggestions, which are rooted in historical narrative reaching back to a period more than a century ago, are ideas now coming to fruition as they form the future research agenda of our domain. This paper encompasses a domain analysis of KO as a means of visualizing the emergence and coherence of our domain, and as a way of denominating the parameters of the universe (or universes) in which our domain operates, as well as the dimensions of the operational paradigms at work. In other words, we look here at the extension and intension of KO as a domain, for the purpose of finding a way to look forward to the 21st century agenda for KO research.

Intensions and Extensions
The parameters, or dimensions, of our domain are given by domain analysis, the representation of a discourse community, in this case of scholars, as can be determined from the formal products, or published research, emanating from the community. From domain analytical studies emerge means for visualizing interaction within the scholarly community, as well as points of shifting emphasis as theoretical paradigms are tested and hypotheses are generated. Knowledge Organization (KO) as a domain is, therefore, no different. Analysis of the domain is relatively easily generated from the formal publications in biennial ISKO conference proceedings, articles in our journal Knowledge Organization, and proceedings of biennial regional chapters, whose meeting years alternate with the ISKO international conferences. According to Tennis (2003) the extension of a domain is the breadth of its topical universe, and the intension is the depth of the paradigms that
comprise it. In this paper I use bibliometric tools to analyze the extension and intension of KO as a domain.

Knowledge Organization is devoted to the conceptual order of knowledge. In the broadest sense KO is the arena in which the heuristics of ordering knowledge are studied. Specifically, KO is the research community devoted to classification and ontology, thesauri and controlled vocabulary, epistemology and warrant, and applied systems for all of the preceding (often, especially in North America, resource description is also considered to be a part of KO). There is a long tradition of the activities and tools of KO; classification, taxonomy, and typology, for example, have always been key to the development of scholarship. Specific applications for information storage and retrieval, such as indexes, bibliographic classifications, etc., have been part of the practice of librarianship and the research agenda of information science since the late 19th century. The formal KO domain, represented by the International Society for Knowledge Organization (ISKO) and its chapters, dates from 1989 (Dahlberg, 2010).

Domain Analysis of Knowledge Organization

Several authors have taken diverse snapshots of KO as a domain over time, such that individual domain analytical pictures are quite segmented. More or less full-scale studies using different source materials and with varying geopolitical emphases are illustrated in Table 1. These studies were compiled for a meta-analytical study reported by Smiraglia (2011c), of which the present paper is an extension. The continuity of analysis of the domain across time and from within its ranks is clear from this table. Over a period of two decades, leading scholars within KO both internationally and regionally, have analyzed the entire breadth of formal scientific publishing in the domain.

Of course, the importance of conference proceedings as a venue for communication in the domain is clear. The limitations of publication in journals are such that only a small proportion of research in the domain appears in journals; most work at the research front emerges in conference proceedings. In the studies shown in Table 1, four journals predominate—Cataloging & Classification Quarterly, Library Resources & Technical Services, The Library Quarterly, and Knowledge Organization. Of these, only Knowledge Organization is exclusively devoted to KO research. In all of these journals readers can find a mix of scientific reports and humanistic narrative analyses. During the period of study these journals would have contained approximately 408 articles; international ISKO conferences contained approximately 550 papers over the same period; American Society for Information Science and Technology/Special Interest Group-Classification Research meetings produced 212 papers. Not counting the contents of regional ISKO chapter conferences, these sources constitute the corpus of formal publication in KO as a domain; most likely the study by Ibekwe-SanJuan and SanJuan (2010) contains the largest portion of the indexed corpus, minus regional conferences.

Citation analysis, author co-citation analysis, and co-word analysis are effective tools for the visualization of the intension and extension of domains. Generating multi-dimensional maps of

<table>
<thead>
<tr>
<th>Citation</th>
<th>Domain</th>
<th>Time</th>
<th>Venues</th>
</tr>
</thead>
<tbody>
<tr>
<td>McIlwaine and Williamson 1999</td>
<td>International trends in subject analysis research</td>
<td>1988-1998</td>
<td>Conference proceedings</td>
</tr>
</tbody>
</table>
domain boundaries helps visualization of a domain’s hypotheses, and thus the direction of its research front (see White and McCain, 1997; Hjørland, 2002; and Tennis, 2003), or as Hjørland might suggest (cf. 2009), the methodologies reveal the socially negotiated constructs at the core of the domain. The studies of KO listed in Table 1 above have revealed meta-level concepts of extension and intension as illustrated in Table 2.

### Table 2: Extension and Intension

<table>
<thead>
<tr>
<th>Domain</th>
<th>Extension</th>
<th>Intension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cataloging &amp; Classification Quarterly 1980-1990</td>
<td>Descriptive cataloging</td>
<td>Management</td>
</tr>
<tr>
<td>Organization of information in The Library Quarterly 1931-2004</td>
<td>Cataloging codes</td>
<td>Costs</td>
</tr>
<tr>
<td>KO 1998-2002</td>
<td>Universal systems</td>
<td>Interoperability</td>
</tr>
<tr>
<td>KO in Spain 1992-2001</td>
<td>Thesauri</td>
<td>UDC</td>
</tr>
<tr>
<td>ISKO 9</td>
<td>Classification</td>
<td>Cognitive and systemic</td>
</tr>
</tbody>
</table>

**Citation**

<table>
<thead>
<tr>
<th>Citation</th>
<th>Domain</th>
<th>Time</th>
<th>Venues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olson 2000</td>
<td>Organization of information in The Library Quarterly</td>
<td>1931-2004</td>
<td>Journal articles</td>
</tr>
<tr>
<td>Smiraglia 2006</td>
<td>ISKO 9 Vienna Cataloging &amp; Classification Quarterly</td>
<td>2006</td>
<td>Conference proceedings</td>
</tr>
<tr>
<td>Roc, Culbertson and Jizba 2007</td>
<td>ISKO in North America Cataloging &amp; Classification Quarterly</td>
<td>2007</td>
<td>Conference proceedings</td>
</tr>
<tr>
<td>Smiraglia 2007</td>
<td>ISKO 10 Montréal Knowledge Organization and Information Organization</td>
<td>2008</td>
<td>Conference proceedings</td>
</tr>
<tr>
<td>Smiraglia 2009</td>
<td>Knowledge Organization in North America</td>
<td>2009</td>
<td>Conference proceedings</td>
</tr>
<tr>
<td>Ibekwe-SanJuan and SanJuan 2010</td>
<td>Knowledge Organization</td>
<td>1988-2008</td>
<td>Journal articles</td>
</tr>
<tr>
<td>Smiraglia 2011a</td>
<td>Knowledge organization in Latin America ISKO 11 Rome</td>
<td>2009</td>
<td>Conference proceedings</td>
</tr>
<tr>
<td>Smiraglia 2011b</td>
<td>Universal systems Resource discovery Thesauri</td>
<td>2010</td>
<td>Conference proceedings</td>
</tr>
</tbody>
</table>
KO shows remarkable coherence as a research domain over time, which is witnessed here by commonality of terminology. The extension of the domain is consistently represented as including theoretical foundations, such as classification and ontology, and epistemology, which lies at the heart of both. The intension is represented by development and testing of applications. The international nature of the domain means that there are occasional overlapping emphases on multicultural and multilingual issues. Geographical diversity does not preclude domain coherence; as Smiraglia (2011b) and López-Huertas and Contreras (2004) suggest, there are domain-coherent parallels among contributors to a Latin American KO conference and a decade of Spanish research. While the extension remains consistent, the intension shifts from universal bibliographic classification to interoperability as the Internet changes scholarship at every level from 1993 onward.

Another indicator of domain coherence is the internal communication within the domain suggested by citation and co-citation of authors, especially those who constitute the research front. Table 3 shows comparative analyses of most-cited authors from four of the studies.

Table 3: Most cited authors in KO

<table>
<thead>
<tr>
<th>I Simposio</th>
<th>NASKO 2007</th>
<th>ISKO 2008</th>
<th>ISKO 2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hjørland</td>
<td>Hjørland</td>
<td>Hjørland</td>
<td>Hjørland</td>
</tr>
<tr>
<td>Lime</td>
<td>Atherton-Cochrane</td>
<td>Neelameghan</td>
<td>Gnoli</td>
</tr>
<tr>
<td>Cabre</td>
<td>Olson</td>
<td>Beghtol</td>
<td>Poli</td>
</tr>
<tr>
<td>Smiraglia</td>
<td>Smiraglia</td>
<td>Buckland</td>
<td>Smiraglia</td>
</tr>
<tr>
<td>Kobashi</td>
<td>Beghtol</td>
<td>Svenonius</td>
<td>Beghtol</td>
</tr>
<tr>
<td>López-Huertas</td>
<td>Beghtol</td>
<td>Kipp</td>
<td>Tennis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Dahlberg</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Vickery</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Broughton</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>La Barre</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>López-Huertas</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mai</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Priss</td>
</tr>
</tbody>
</table>

Table 1 Hjørland and Beghtol are consistently among the most-cited; together with Dahlberg, and Vickery, they could be said to constitute the theoretical core of the domain—its extension, in other words. The other most-cited authors vary from study to study, and seem in each case to be aligned with thematic emphases of the domain’s intension, as it varies from year to year or place to place. For instance, terminology is an emphasis of the Mexican conference in the first column, facets are emphasized in NASKO 07 and ISKO 11, and so forth.

Author co-citation analysis maps help further to visualize these results. Figure 2 contains MDS maps from 2007s I Simposio in Mexico City, 2008s International ISKO conference in Montréal, and 2010s International ISKO conference in Rome.

In every case, Hjørland anchors the visualization at the far right, and in all three cases one of two large clusters, representing theoretical foundations, wraps around his name. Citing his work is essential to the entire domain. In all three visualizations, the second large cluster represents both the intension and the emergent research front. In the Latin American conference the emphasis is on terminology, language, controlled vocabulary, and thesauri. In Montréal 2008, the second cluster stretches from Jörgesson to Renear but is anchored internally by Olson, Green, and Greenberg, and is associated with Foucault. In Rome 2010 the second cluster envelopes Vickery, Beghtol and the research front. In all three cases we see a dynamic tension between KO as a domain and KO systems, between theory and applied, and between humanistic and scientific or technological approaches.
In 2011 and 2012 Knowledge Organization has featured the 3-5 best papers of each regional conference; typically these papers are peer-reviewed before the conference, then selected by conference chairs, and final versions are submitted to KO. To close this analysis, let us look at the keywords in the titles of those papers (see KO for full texts).

Table 4: Themes from 2011 Regional ISKO Conferences

<table>
<thead>
<tr>
<th>NASKO</th>
<th>Morocco</th>
<th>Spain</th>
<th>Italy</th>
<th>Brazil</th>
<th>France</th>
</tr>
</thead>
<tbody>
<tr>
<td>21st century</td>
<td>Agricultural research in</td>
<td>Authority control</td>
<td>Colon</td>
<td>Documentary systems</td>
<td>Activities</td>
</tr>
<tr>
<td>Categorizing</td>
<td>Algeria</td>
<td>Complex systems</td>
<td>Classification</td>
<td>systems</td>
<td>awareness</td>
</tr>
<tr>
<td>Classification</td>
<td>Classifications</td>
<td>Documentation</td>
<td>Indexing and</td>
<td>scientifica</td>
<td>Authorial</td>
</tr>
<tr>
<td>DDC</td>
<td>Linked open data</td>
<td>Global agents</td>
<td>classification</td>
<td>production</td>
<td>stance</td>
</tr>
<tr>
<td></td>
<td>data</td>
<td>Document KOS</td>
<td>systems</td>
<td>Catalog</td>
<td></td>
</tr>
<tr>
<td></td>
<td>mapping</td>
<td>indexing Paradigms</td>
<td>Knowledge</td>
<td>Pragmatism</td>
<td></td>
</tr>
<tr>
<td></td>
<td>People-centered</td>
<td>Global</td>
<td>standards</td>
<td>Documental</td>
<td></td>
</tr>
<tr>
<td></td>
<td>properties</td>
<td>catastrophic risks</td>
<td>Multimedia</td>
<td>dimensions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Prototype</td>
<td>IR in Arabic</td>
<td>information</td>
<td>Epistemic</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Knowledge</td>
<td>retrieval</td>
<td>Faceted</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nuovo</td>
<td></td>
<td>classification</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ontic</td>
</tr>
</tbody>
</table>
These papers, then, represent the cutting edge in KO research in 2011, in advance of the 2012 International ISKO conference in India. Look at the breadth of intension represented here, mostly in applied systems developments, and all directed in one way or another toward the domain’s extension of interoperability. We also see, again, the dynamic tension between humanistic enterprises (e.g., prototype theory) and scientific approaches (e.g., query reformulation).

Meta-Analysis Suggests Dimensions

Taking together the evidence outlined here, we see that KO as a domain demonstrates coherence across time and across geo-political boundaries, particularly as it concerns its theoretical foundations. KO as a domain has robust and continuous formal publication venues that help to maintain domain coherence. In KO, theoretical poles are both conceptual and methodological. The domain is scientific, but also has deep roots in humanistic methods and modes of thought. Differences that emerge in intension reflect shifting cultural approaches across regions and across time. We observe a shift over time in intension as the domain moves from emphasis on universal classifications to interoperability. Thus we see consistently marked dimensions within the domain: theoretical versus applied on one continuum, humanistic versus scientific on another. These dimensions serve to maintain constructive and dynamic tension within the domain, which in turn keeps the research front constantly in a state of renewal. Problems of classification theory and KOS relate to IR, but with the dynamic tension between humanistic and empirical approaches; domain analysis in KO is a tool for developing KOS. Multiple dimensions, then, in dynamic tension, continually push the boundaries of the theoretical and scientific/applied paradigms. Faced with different universes of knowledge, KO as a domain approaches the problems of analysis and concept designation from within the dynamic tension demonstrated here. Collins (1998) suggests no school of thought can comprise more than six points of view without either concretizing or splitting. KO itself has concretized, maintaining its status as a domain by using this tension to rein in emergent trends that would threaten explosive expansion—e.g., social tagging, folksonomy.

References


Birger Hjørland - Royal School of Library & Information Science, Denmark

Knowledge Organization = Information Organization?

Abstract:
Are the terms “information organization” (IO), “organization of information” (OI) and “information architecture” (IA) synonyms for knowledge organization (KO)? This study uses bibliometric methods, among others, to determine some relations between these terms and their meanings. Apparently the data shows that these terms should not be considered synonyms because each of the terms IO, OI, IA and KO produce a different set of high ranked authors, journals and papers. In many cases the terms are, however, used interchangeably (and thus indicating synonymity) and it is argued that the underlying theoretical principles are identical but that the different terms tend to be applied in different contexts: KO in the library context, IA in the web-context and IO and OI in more unspecified ways.

Introduction
The present study is concerned with the relations between four terms from the literature of library and information science (LIS):
- Information organization (IO),
- Organization of information (OI),
- Information architecture (IA) and
- Knowledge organization (KO)

More precisely, it is about whether or not these terms should be considered synonyms? Synonymy being defined as the semantic relation that holds between two terms that can—in a given context—be said to express the same meaning. The term KO is well established and the International Society for Knowledge Organization (ISKO) and its publications, including the journal Knowledge Organization, are among the core actors in this field. IA, on the other hand, is a rather new term, which in some contexts seems to be more “hot,” technologically advanced or prestigious term. The two other terms: IO and OI are included in this examination in order to clarify the meaning of closely related terms. Are there differences in meaning or are the different expressions attributable, in part, to what Konrad (2007) termed “por terminological hygiene”? The methodology applied in this study is also suggested for examining concepts and relations in other fields and it is therefore an approach to KO applied on the field itself.

Method
Each of these four terms were searched in Social Science Citation Index (SSCI) both in the whole database (Table 1) and limited to LIS (Table 2) in January 2011. For each term and each database was ranked 1) the most cited authors 2) the most cited journals or works and 3) the most cited references. The content in these tables is analyzed. Core texts in KO and IA are also examined in order to compare the theoretical issues involved.

Results
The rankings of the bibliometric investigation are displayed below.
Table 1: Top 5 Rankings of authors, works and papers in Knowledge Organization and Information Organization in SocialSciSearch, all subject fields (January 2011)

<table>
<thead>
<tr>
<th>Most cited authors</th>
<th>Knowledge Organization, KO</th>
<th>Information Organization, IO</th>
<th>Organization of Information, OI</th>
<th>Information Architecture, IA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank #1</td>
<td>Hjorland B</td>
<td>Miller Ga</td>
<td>Duncan J</td>
<td>Rosenfeld L</td>
</tr>
<tr>
<td></td>
<td>Dahlberg I</td>
<td>Svenonius E</td>
<td>Williamson Oe</td>
<td>Nielsen J</td>
</tr>
<tr>
<td></td>
<td>Beghtol C</td>
<td>Baddeley A</td>
<td>Alchian Aa</td>
<td>Brancheau Jc</td>
</tr>
<tr>
<td></td>
<td>Chi Mth</td>
<td>Porter Me</td>
<td>Posner Mi</td>
<td>Wurman Rs</td>
</tr>
<tr>
<td></td>
<td>Kogut B</td>
<td>Zand De</td>
<td>Kahneman D</td>
<td>Marchionini G</td>
</tr>
<tr>
<td>Rank #2</td>
<td>J Doc</td>
<td>Psychol Rev</td>
<td>Psychol Rev</td>
<td>Information Architect</td>
</tr>
<tr>
<td></td>
<td>Knowl Organ</td>
<td>J Am Soc Inform</td>
<td>Cognitive Psychol</td>
<td>Commun Acm</td>
</tr>
<tr>
<td></td>
<td>Cognitive Psychol</td>
<td>Cognitive Psychol</td>
<td>Percept Psychophys</td>
<td>Sci</td>
</tr>
<tr>
<td></td>
<td>Acad Manage Rev</td>
<td>J Exp Psychol Learn</td>
<td></td>
<td>Architecture</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rank #6</td>
<td>Most cited references</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Top 5 Rankings of authors, works and papers in Knowledge Organization and Information Organization in SocialSciSearch, Information and Library Science (January 2011)

<table>
<thead>
<tr>
<th>Most cited authors</th>
<th>Knowledge Organization</th>
<th>Information Organization</th>
<th>Organization of Information</th>
<th>Information Architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank #13</td>
<td>Hjorland B</td>
<td>Svenonius E</td>
<td>Case Do</td>
<td>Rosenfeld L</td>
</tr>
<tr>
<td></td>
<td>Dahlberg I</td>
<td>Belkin Nj</td>
<td>Davenport Th</td>
<td>Nielsen J</td>
</tr>
<tr>
<td></td>
<td>Beghtol C</td>
<td>Choo Cw</td>
<td>Fidel R</td>
<td>Brancheau Jc</td>
</tr>
<tr>
<td></td>
<td>Ranganathan Sr</td>
<td>Ingwersen P</td>
<td>Kwasnik Bh</td>
<td>Dillon A</td>
</tr>
<tr>
<td></td>
<td>Soergel D</td>
<td>Taylor Ag</td>
<td>Paton Mq</td>
<td>Marchionini G</td>
</tr>
</tbody>
</table>
Knowledge Organization

Information Organization

Organization of Information

Information Architecture

<table>
<thead>
<tr>
<th>Most cited journals/works</th>
<th>Most cited references</th>
</tr>
</thead>
</table>

Data Analysis

If we compare the five most cited authors in the whole of SSCI the first observation is that there is no overlap: Each of the four concepts has a unique set of most cited authors (which of course change if more than just the top five is considered; data not shown), seemingly indicating that we are dealing with four separate fields.

a) The term KO is dominated by authors from LIS: The three most cited researchers in rank #1 often attend the same conferences and publish in the same journals. Dahlberg is the founder of the International Society for Knowledge Organization (ISKO) and the journal Knowledge Organization. In rank #1 only Chi et al. (1981) and Kogut & Zander (1992) are from outside LIS (respectively from cognitive science and knowledge management) indicating that the term is also used in those fields. The first three are well known researchers in LIS as are all researchers in rank #13. The three most cited journals are from LIS: J.Doc, Knowledge Organization and JASIST. The fourth and fifth most cited journals are from psychology (Cognitive Psychology) and Management (The Academy of Management Review).

b) Concerning the term IO: Cognitive psychologist G.A. Miller was most cited in rank #4 (however in a new search made on 2012-01-02 Svenonius and Miller switched places). Elaine Svenonius is a well know scholar in KO. Her book The Intellectual Foundation of Information Organization (2000) is clearly a work from the tradition of LIS and KO, which has chosen the label IO rather than KO (and therefore indicating synonymity between these terms). Her book appears under both KO (rank 15) and IO (rank 6+18). Returning to rank #4: A. Baddeley is a cognitive psychologist, while M.E. Porter and D.E.Z and are management scholars. In the social sciences, the term IO is thus not dominated by LIS.
researchers, and within LIS (rank #16) Nicolas Belkin and Peter Ingwersen are not foremost known for their contributions to KO. C.W. Choo is researcher in knowledge management. Arlene G. Taylor is a well-known author of a textbook in KO (Taylor & Joudrey, 2009). It appears that IO is sometimes used as synonym for KO, but in general what is found under that term is very mixed.

c) The term OI is mostly used by psychologists and cognitive scientists (whether or not cognitive science is a fruitful theoretical basis for KO cannot be discussed in the present paper). In Rank #7 none of the researchers are from LIS. In rank #19 Raya Fidel and Barbara Kwasnik are from KO, the others from other subfields of LIS. This term is therefore the term with the weakest link to KO and what is found under it is also very mixed.

d) Finally, the term IA designates what appears to be a “new” field. A core text is Morville & Rosenfeld (2006) (1st edition: Rosenfeld & Morville, 1998) and this text is the highest ranking in both rank #11 and #23 just as one of the authors is the highest ranking author in both rank #10 and #22. Although this book is focusing on web-design, it contains a lot of traditional topics from KO, such as hierarchy, folksonomies, metadata, thesauri, and faceted classification. I am not saying that nothing is new in this field (and it is certainly attracting some talented people), but I would say that it is an exaggeration to speak of a new field because the overlap with KO is high, and the intellectual basis is too closely related (in other words: each field is too small in substantive content to be separated from the other). In my opinion IA is to some extent “old wine in new bottle” and the tendency to create new labels may have some negative effects in fragmenting the field.

Discussion

Does KO = IO? What differences does it make whether we prefer the term knowledge or the term information in LIS and in KO? There are different views on this issue in the literature. D. A. Kemp (1988, p.3) argued that “knowledge retrieval” should substitute “information retrieval”; Van Rijsbergen and Lamas, on the other side, wrote:

— In the early days of Information Retrieval (van Rijsbergen, 1979), people used to qualify their statements about information retrieval (IR) by saying that really they were working on document retrieval. It was denied strenuously that information was being retrieved. As Lancaster (1968) wrote, “An information retrieval system does not inform (i.e., change the knowledge of) the user on the subject of his inquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request.”

The situation has changed. We believe that the purpose of an information retrieval system is to provide information about a request and that a request is a representation of an information need that an IR system attempts to satisfy.” (van Rijsbergen & Lalmas, 1996, p. 386).

There are strong indications that the term ‘information’ became popular with library science and documentation more because of its appeal than for its scientific merits (cf. Capurro & Hjørland, 2003; Hjørland, 2000; Furner, 2004). These authors, among others, argue against van Rijsbergen & Lamas’ point of view. A document can be said to materialize the knowledge produced and thus to represent knowledge. Documents may also be said to have the potential to inform people. The criteria of when documents represent knowledge (what is knowledge?) or when documents inform people (what is information?) have been the focus of much discussion. Buckland (2012), for example, finds that information science is concerned with what people know (i.e., with knowledge), and his arguments are related to a deeper concern about the fruitful development of LIS: it is rather important issues that are at stake. It may be argued that knowledge and information can be used as synonyms in LIS, and a textbook by Rowley & Hartley (2008) used the title Organizing knowledge but adds the subtitle: An Introduction to Managing Access to Information. In this way some authors may try to attract people whichever of these terms
they might prefer and again indicating the connection between the terms IO, OI, IA and KO. I’ll argue, however, that knowledge should be the preferred term in LIS—and thus that KO should be preferred among the four terms considered in this article.

The present study has used bibliometric methods and has considered different disciplines, which is a concept in the sociology of science. The methods and theories used here are thus much more related to fields like the theory of knowledge and the sociology of knowledge than to information theory, indicating an important relation to other disciplines concerned with knowledge. My suggestion is, in other words, that the term knowledge moves us relatively away from fields like information theory and computer science towards fields such as social semiotics, science studies and the study of documents and their role in human activities (activity theory). I believe that such a social turn is very important for developing LIS as a scholarly discipline.

Subject terminology should not be used as buzzwords. There is a tendency to change terminology in this way. Sheila Webber shows how many courses in England shift titles from information science to information management simply because the word science does not have the same appeal among students trying to choose an education programme while the term "management" is popular among students. She further puts the question (p. 328): ""Library and Information Management": is it merely an umbrella term and administrative convenience? Is it a new name for IS? Is it a different discipline?" In a similar way may many phrases containing the word information (i.e. information retrieval, information organization, etc.) are being used more because of their appeal than their scientific merits. At the School of Information Studies at the University of Wisconsin-Milwaukee, for example, the name of the subject has recently changed from KO to IO. This is just a change in name, not a change in what is being taught. I do not believe that it is healthy for scholarship to use terminology as buzz-words in order to attract students, to try to raise the image of a dusty profession, to follow fad or whatever. I do not believe that science and scholarship should be constructed on the basis of what can be sold. It is the other way round: Things should be sold because they have inherent qualities, which the broader society learns to respect and in this way making the names of the fields popular.

An analysis of the theoretical problems involved demonstrates that all of the fields: KO, IO, OI and IA are primarily concerned with subjects, concepts, and semantic relations between concepts. The basic theoretical knowledge is therefore the same in the fields covered by the four terms, although IA is more about organizing subjects and concepts on the web, whereas KO has traditionally been more (but not exclusively) related to libraries and bibliographic databases. However, from the perspective of academic research, such differences are superficial, not essential.

In a thesaurus for the domain of LIS, Knowledge organization (KO) should therefore be the preferred term (descriptor), while the other examined terms: IO, OI and IA should be lead-in terms (also termed non-preferred terms, synonyms, non-descriptors or entry terms).
Conclusion

This study has argued that it might be a good idea to continue to use the term *knowledge organization* and to connect KO better with other disciplines devoted to the study of knowledge.

The study has also explored the contextual issues related to the use of the four terms KO, IO, OI, and AI. Philosopher Wittgenstein is famous for his “use theory” of meaning: You have to study the use of language in order to understand its meanings. Miller & Leacock (2000) raised the following question: “Why isn’t a dictionary a good theory of the lexical component of language?” The answer they provide is that dictionaries lack contextual information that would enable a user to make the correct association between senses and actual contexts. They provide the example, *Our families erode a lot*, which sounds bizarre until you read the definition of *erode*: “eat out, eat away”. Thesauri—and most kinds of knowledge organization systems (KOS)—also lack such contextual knowledge (this is not, however, the case with, for example, historical dictionaries which may provide detailed information about how words have been used).

The shortcoming of traditional KOS may be countered by bibliometric studies such as the one made in the present article: this is a way to examine the terms in different contexts in which their meanings are negotiated and may be more or less stabilized (KO and IA seem rather stabilized compared to IO & OI). The study has thus demonstrated how bibliometrics accompanied with a study of the contents of the most cited works may be used in order to study how concepts are used in different fields and thereby as a tool for organizing knowledge.

References


H. Peter Ohly – ISKO, Germany

Mission, Programs, and Challenges of Knowledge Organization

Abstract
Traditionally “Knowledge Organization” is the ordering of documents on a bookshelf and the indexing of these in a more or less one-dimensional catalog. Now with the Web, its many and hyperlinked, distributed and heterogeneous sources the plain terminological approaches are no longer sufficient. In the same way ISKO society has to rethink its mission.

The “New” Knowledge Organization, which aims at the “Semantic” Web 3.0 tries to combine different Knowledge Organization systems by shared meta data and formalized ontologies that are logical deductive. Whereas the semantic approaches have quite an opposite approach to the user driven systems the future might lay in a combination of logical descriptions and self-organizing principles, what could be named “Self Organizing Knowledge Organization Systems”. But still there are open questions which might be solved in future by ISKO and neighbour societies.

What is Knowledge Organization?
Dahlberg (2006) defines Knowledge Organization as “… the science of structuring and systematically arranging of knowledge units (concepts) according to their inherent knowledge elements (characteristics) and the application of concepts and classes of concepts ordered by this way for the assignment of the worthwhile contents of referents (objects/subjects) of all kinds”. More precisely, Dahlberg (1998) defines knowledge organization as:

A subject area encompassing the organizing of a) units of knowledge concepts and b) all types of objects (minerals, plants, animals, documents, pictures, museum objects, etc.), related to particular terms or categories, so as to capture what is known about the world in some orderly form allowing it to be further shared with others.

Knowledge organization encompasses the following nine sub-areas:
1. the epistemological, mathematical, system-theoretical, cognitive scientific and scientific theoretical premises of order of concepts as well as their historical background,
2. the knowledge of elements and structures of systems of concepts,
3. the methodology of intellectual construction, conservation and revision of this system and computerization; including questions of paradigmatic and syntactic relating of their elements and units as well as keeping the system compatible and evaluating this system,
4. the methodology of intellectual and machine applications of this system via classification and indexing,
5. the knowledge of existing universals and
6. special taxonomies and classification systems including documentation language (thesauri),
7. questions arising from the influential areas linguistics (~ linguistics mathematics) and terminology; including the retrieval problems, especially in online access,
8. the application of content indexing of all types of documents and in all subject areas,
9. the entire periphery of knowledge organization in the workplace, individual centers, societies, countries and in international areas, as well as the question of education, the economy, the user, etc.
With respect to e.g. social tagging, the systematic approach and the presumption of inherence of the knowledge elements are nowadays no longer obvious. Traditionally “Knowledge Organization” was understood as the ordering of documents on a bookshelf and the indexing of these in a more or less one-dimensional catalog. At least with more detailed documents (e.g. journal articles), the growth of data bases and the availability of sophisticated retrieval techniques, the concept approaches so far were not satisfactory and term based organization became increasingly important. Now with the Web, its many and hyperlinked, distributed and heterogeneous sources the plain terminological approaches are no longer sufficient. Zeng (2008) sees an increase of dimensionality and functions with the development from term lists, via metadata-like models and classifications, to relationship models (see Fig. 1). In principle this reflects a shift from more subjective views of knowledge and its order to more objective ones. Hence it has to be rethought how far e.g. ISKO (International Society for Knowledge Organization) has to change its mission.

There are many disciplines which are partially applied in Knowledge Organization or which are operating in close connection with it, e.g. Linguistics, Cognitive Science, Philosophy. As discussed above Knowledge Organization is more oriented to the labeling, arranging, and retrieval of knowledge in archived documents. Main frontiers exist nowadays with Knowledge Management and Knowledge Engineering (cf. Fig. 2). In Knowledge Management the main focus is to make profit out of the knowledge within an enterprise and its business. The codification and retrieval of knowledge for management processes are part of it. Knowledge Engineering tries to mechanize knowledge storage, its maintenance, and its integration. Human models of understanding the universe of knowledge are part of it. Whereas Knowledge Management and Knowledge Engineering are appropriate to shape the primary scope of Knowledge Organization, these are also neighbor disciplines worth to be considered as essential collaboration fields. E.g. “enhanced” and “electronic” hyphened application areas are more than a sum of all of these but a true progress if combined sophisticated, like intended in the Semantic Web.
Challenges for Knowledge Organization

At several conferences the future of Knowledge Organization was discussed, 2006 in the Vienna ISKO conference (Ohly 2008a) Winfried Schmitz-Esser realized that a world model is needed that integrates means of organization and that can also detect and process knowledge in texts. And Gerhard Budin stated that Knowledge Organization nowadays comprises cognitive, epistemic, communicative, and automatic knowledge representation, creation, and processing. On the IKONE conference 2007 in Bangalore Maxmilian Stempfhuber stressed out that aggregation of data resources in portals requires special treating of heterogeneity with respect to user demands. One of the most demanding statements given at this conference was by Prasad Bhaarat Ram: “Give me what I want, not what I ask for” (Ohly 2008b). Means for that might be procedures that deal with misspellings, spam ranking, user models. A conference on „Scientific Communication of the Future” in Jülich 2007 yielded in diversification of knowledge and types of knowledge communication distinguishing between mainstream knowledge and ingenious knowledge creation.

These discussions of the last decade shape clearly, that the users as well as the indexers have a fluid notion of knowledge and document contents to be useful for a certain application. These users are very different with respect to their individual or public tasks and their observing or constructing role, especially with the “Social” Web 2.0 (Trump 2007; see Fig. 3). Whereas in the pre-digital world the experts were dictating the values to the users, now in a native digital world the public and skilled users are setting the norms for the scientific experts (cf. Quoniam 2009). The information quality is in so far not static but must be able to adopt information sources from yesterday to arising questions of tomorrow (see Fig. 4). Prerequisite - though not sufficient - is the trust in the information creation and its creating institutions. The retrieval should apply sophisticated robust processing techniques which are able to mine important information from many and heterogeneous data bases. In favor of the current user a flexible adaption with selective narrative ranges for decision making have to be provided for. Hence the demanding question will be: How can outdated information be transformed and reused for future problems. Nevertheless unobvious user expectations can hardly be matched by a mechanical information system. At more recent conferences (ISKO France 2009, Lyon and 2011, Lille; ISKO Germany 2009, Bonn; UDC Seminar 2011, Den Haag; DGI 2012, Düsseldorf) had similar discussion rounds but picked out more the lacking of Web semantics, the conflict with new social media and the computer science dominance 3.

---

1 International Conference on Future of Knowledge Organization in the Networked Environment
2 Ball 2007
One key question is the theoretical foundation of information and Knowledge Organization. If we are going back to Dahlberg’s definition, we are finding also questions of application and periphery of Knowledge Organization. And the founders of classification systems, like Dewey or Otlet (Rayward 2010), went far beyond the question of positioning and distributing books. Such a focus on usability is accompanied by questions like: What kind of knowledge is worth being collected? Where should we get it from? How can we use and preserve it? Who should make profit out of it? There are many theories offered from Activity Theory to Design Theory, not to omit theories that include ecological and global aspects. A look only on Complexity Theory (Morin 2006) shows the wide range of theories nested with it under different aspects (see Fig. 5): System Theory (Bertalanffy) and Cybernetics (Wiener) up to Global Network Society (Castells). Thus Knowledge Organization must be aware of different theoretical positions that enforce certain tasks and visions and must be able to lay its emphasis and principles on those theories that are appropriate in its special application area and for its user clientele. Certain

---

Fig. 3: Web 2.0 Users Typology (translated from Trump 2007)

Fig. 4: Past-Present–Future of Information (Ohly 2011b)

Theoretical Foundations of Knowledge Organization

One key question is the theoretical foundation of information and Knowledge Organization. If we are going back to Dahlberg’s definition, we are finding also questions of application and periphery of Knowledge Organization. And the founders of classification systems, like Dewey or Otlet (Rayward 2010), went far beyond the question of positioning and distributing books. Such a focus on usability is accompanied by questions like: What kind of knowledge is worth being collected? Where should we get it from? How can we use and preserve it? Who should make profit out of it? There are many theories offered from Activity Theory to Design Theory, not to omit theories that include ecological and global aspects. A look only on Complexity Theory (Morin 2006) shows the wide range of theories nested with it under different aspects (see Fig. 5): System Theory (Bertalanffy) and Cybernetics (Wiener) up to Global Network Society (Castells). Thus Knowledge Organization must be aware of different theoretical positions that enforce certain tasks and visions and must be able to lay its emphasis and principles on those theories that are appropriate in its special application area and for its user clientele. Certain

---

4 See also: Montuori 2004
5 Bertalanffy 1957; Wiener 1948; Castells 2005
positions determine the kind of offering information as well as expectations of the users. Not at least psychological theories, how we are attracted to use information (Berlyne: Activation Theory) and which design of information will be fashionable Bürdeck: Design Theory), as well as sociological theories, how information and information norms are diffusing (Rogers: Diffusion Theory) and psycho-social dynamics of opinion leadership (Lewin: Group Dynamics) have to be considered. Hjørland (2002; see: Fig. 6) distinguishes between Empiricism (induction from data), Rationalism (logical modeling), Historicism (historical background), and Pragmatism (goals and values of information and users). Hjørland (1997) explains the “Pragmatic Theory of Knowledge” among others as follows:

Fig. 5: Complexity Map (from Castellani 2009)

---

6 Berlyne 1967 (see also: Scott1966); Bürdek 2005; Rogers 1962; Lewin 1935
7 Extract from Hjørland (1997): <http://www.iva.dk/jnl/lifeboat_old/Positions/Pragmatism.htm>
8 Under which he subsumes his information concept of Activity Theory (Leontev 1978)
“Since living and acting constitutes the a priori of knowledge, knowledge is constructed in such a way that an application of well constructed knowledge will directly or indirectly serve living and acting” […] “in a bio-physical, a socio-cultural and a subjective world”. […] “There is a continuous interaction between knowledge and action so that knowledge is created in and through action and so that experiences that the actor acquires through action influences subsequent action.”

**New Knowledge Organization**

The “New” Knowledge Organization, which aims at the “Semantic” Web 3.0 tries to combine different Knowledge Organization systems in the Internet by shared meta data and formalized ontologies which are logical deductive. Examples for these are the NKOS group with recommendations for thesauri description schemes (NKOS 1998) and SKOS which applies machine-readable resource description languages to knowledge organization systems such as thesauri, classification schemes, and other concept schemes (Miles/Bechhofer 2008). The SKOS data model provides basic sets of documentation properties, semantic relation properties, lexical labeling, label relations, concept collections, and concept mapping. These sets provide a framework that can be adapted to more specific needs.

Whereas the semantic approaches have quite an opposite approach to the user driven “social” systems 10 the future might exist in a combination of logical descriptions and self-organizing principles, what could be labeled “Self Organizing Knowledge Organization System” (SOKOS). A first attempt in harmonizing Knowledge Organization systems with a machine-readable world is given by the ISO norm 25964 Information and Documentation, part 1 and 2 (2011/2012) 11, that revises the existing international standards for retrieval adequate thesauri and is also intended for their interoperability with other vocabularies in a Semantic Web context.

---

9 Order changed by the author
10 See e.g. Flickr <http://www.flickr.com/>
11 See also: Dextre Clarke 2011 and 2011/2012

---

<table>
<thead>
<tr>
<th>Empiricism</th>
<th>Rationalism</th>
<th>Historicism</th>
<th>Pragmatism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevant:</td>
<td>Relevant:</td>
<td>Relevant:</td>
<td>Relevant:</td>
</tr>
<tr>
<td>Observations, sense data.</td>
<td>Thinking, logic, mathematical</td>
<td>Background, knowledge about</td>
<td>Information about</td>
</tr>
<tr>
<td>Induction from collections of</td>
<td>modeling, systems of axioms,</td>
<td>pre-understanding, theories</td>
<td>goals and values and</td>
</tr>
<tr>
<td>observational data.</td>
<td>definitions and theorems.</td>
<td>connections.</td>
<td>consequences both</td>
</tr>
<tr>
<td>Intersubjectively controlled</td>
<td>Low priority is given to empirical data because such data must be organized in accordance with principles which cannot come from experience.</td>
<td>Intersubjectively controlled data are often seen as trivia.</td>
<td>involving the researcher and the object of research (subject and object).</td>
</tr>
<tr>
<td>data.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-relevant:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speculations, knowledge</td>
<td>Low priority is given to empirical data because such data must be organized in accordance with principles which cannot come from experience.</td>
<td>Low priority is given to decontextualized data of which the meanings cannot be interpreted. Intersubjectively controlled data are often seen as trivia.</td>
<td>Low priority (or outright suspicion) is given to claimed value-free or neutral information. For example, feminist epistemology is suspicious about the neutrality of information produced in a male dominated society.</td>
</tr>
<tr>
<td>transmitted from authorities.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Book knowledge&quot; (&quot;reading nature, not books&quot;). Data about the observers' assumptions and pre-understanding.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Knowledge Organization has to turn back to formal, semantic approaches, such as facetted concept systems. But as well it has to be logically more precise and enable reasoning over multiple sources in the Web environment. But also self-adapting approaches, such as social indexing and its quantitative exploitation can be neglected no longer.

Conclusions for ISKO

ISKO as a specialized society in this field has to help its clientele to get orientation in an unclear offer of competing approaches and specific realizations. Means for this are international and interdisciplinary congresses, articles in its specialized journal (Knowledge Organization), access to classical texts, text books, repositories, and exchange relations with other societies (cf. ISKO 1989, Art. 4). Not at least the theoretical background of information, resp. knowledge, and its coding and exchange must be refurbished, especially it has to include the social dimension. A step forward into this direction are repositories with classical readings and old Knowledge Organization journal issues, as well as a specialized bibliography open to the community and the encouraging of norms and evaluations concerning the field of Knowledge Organization. Knowledge Organization textbooks have to include the question of ontologies, machine processing, and Web environment. Exchange with other disciplines or even special groups requires more active contacts with other scientific societies in the field as well as clear exchange rules with these. Nevertheless ISKO has to define its boundaries to other sciences.

But there are still open questions that might be solved in future by ISKO and neighbor societies:

1. What are indicators for “good” Knowledge Organization System?
2. How can Knowledge Organization System with different principles (facetted, social, etc.) be combined?
3. How can local heterogeneity be combined with the aim of a global e-science?
4. Who is the target for Knowledge Organization literacy affords?
5. What is the profile of a Knowledge Organization profession?
6. Will Knowledge Organization only become a mere application of Knowledge Engineering in the field of information brokering?

Outlook

Knowledge Organization is an everlasting demand for efficient storage, retrieval, and processing of information codified in text and other human artifacts. Similar to Information Science it is such common that it is sometimes not seen as an own discipline with the prerequisite of getting standardized, established, mediated, and updated. With the development of more and more complex information and communication technologies it became increasingly dependent from these and is sometimes understood as sub-discipline of these. Without any intention to subordinate one to the other it must be clearly seen that engineering means and the semantic cultural logic are different aspects of one more or less application problem. The problem is to provide knowledge of a special kind for a special purpose (cf. knowledge management in enterprises). These situations should be closer defined by Knowledge Organization and then knowledge engineering techniques can be applied in cooperation. On the other hand Knowledge Engineering is necessary for the proliferation of adequate techniques for Knowledge Organization in whatever context if the intention is clearly enough specified. Currently we have the offering of available techniques without knowing how appropriate they are for application problems. E.g. the combination between open and closed techniques is a question for knowledge workers according to its social functionality and desirability. It is up to the Knowledge Organization to provide a transfer from the one to other. In so far information technology (now computer based) is a prerequisite for every information designer but not the only and not the main focus. To find
its role it is necessary to explore the philosophical background for the current knowledge intermediation.

References


---

12 All Internet links visited April 28th 2012


The Basic Concepts Classification

Abstract
This paper describes the Basic Concepts Classification. It discusses its motivation, advantages, novel structure, and feasibility.

This paper summarizes the Basic Concepts Classification (hereafter BCC), and describes its advantages. Though I have provided a philosophical justification for this classification elsewhere (2003a, 2008, 2011a) and outlined key aspects of it elsewhere (2003b, 2011a, 2012), I have not previously outlined the classification in its entirety.

The Basic Concepts Classification takes its name from an argument that complex concepts which convey quite different meanings across disciplines and cultures can be broken into basic concepts for which the degree of shared understanding across groups is (at least potentially) great enough to support a universal classification of documents. Critically, basic concepts are most feasible with respect to the phenomena of the natural and social worlds and the relationships that exist among them (Szostak 2011a).

Justification
Why a new classification? Simply put, scholars, and especially interdisciplinary scholars, are poorly served by existing classifications. As the Leon Manifesto (see www.iskoi.org/ilc/leon.htm) has noted, scholars want to search by causal relationship (that is, X is alleged to influence Y), theory applied, and method applied. The first is difficult at present, the second and third virtually impossible. The proposed classification will facilitate recovery by both scholars and general users.

Uses
Though designed as a stand-alone classification, the BCC might also be used as a supplement to an existing scheme. Empirical work reported in Szostak (2011a and 2011b) shows how entries in the Dewey Decimal Classification (DDC) can be readily translated into BCC, and their meaning clarified in the process. The BCC could provide a logical structure on which to make additions to the DDC or other classifications.

Structure
The essence of the classification is:
1. Classification in terms of phenomena and relationships rather than disciplines
2. Phenomena are disaggregated usually by „type of,” but occasionally also/instead by „elements of”
3. The disaggregation follows a logical deductive format – grounded in an ontological understanding of the world – but this is supplemented by an inductive „literary warrant” approach – itself informed by an epistemological appreciation of the academy – that ensures that all topics are represented.
4. A universal classification of „properties (qualities)” has also been developed to clarify phenomena and relationships. Compounding can thus be employed for „red house.” Space can be saved in the schedules also by compounding „talk” plus „loudly” for „shout”.
5. Classification of most works as relationships between phenomena. That is, linked notation is used to a much greater extent than in any other classification. This reflects
6. an epistemological understanding that most research investigates relationships (especially causal) among phenomena. Linked notation is also used for „properties” of a thing. The result is a classification that not only conserves on notation but is much more useful to the user.1

7. Relationships themselves are classified using a universal classification of relators developed (using a mix of deduction and induction to ensure a logical structure but also exhaustiveness) in Szostak (2012). Notably, compounding of relators (with other relators or with phenomena or qualities) allows hundreds or even thousands of distinct types of relators to be indicated through recourse to a classification of less than one hundred basic relators.

8. Use of expressive notation as much as possible.2

9. Use of short notations for simple subjects. In this way compound notation becomes feasible even for very complex subjects.3

10. Use of compound notation. Note, though, that commonly used compounds can be given a simpler notation if necessary.

11. Cutter numbers are used whenever there are long lists of types of a well-defined class (authors, countries, theories). This is especially useful when the list expands through time.

12. Detailed classifications of methods and theory types so that works can be precisely classified in terms of the theories and methods applied as well as the phenomena studied. Scholars often care not just (or primarily) about what a work is about but what theories and methods were applied. It is at present generally impossible to search by theory or method applied.

The classification takes a „faceted” approach as recommended by almost all contemporary theorists of classification. Yet it appreciates that different facets need to be treated in quite different ways. In particular, it distinguishes facets that apply to a particular phenomenon from facets that apply to a relationship among phenomena. Given the importance of facet analysis, it is useful to relate here how the 13 facets identified within the Bliss Classification are dealt with in this classification. As in Bliss, the treatment of different facets in different ways allows us to eschew the use of dedicated facet indicators:

1) Thing. While Bliss treats „Thing” as the principal foci of interest in a discipline, the BCC is organized around real phenomena that exist in the world (and can thus at least potentially be studied by multiple disciplines). As with Bliss and other classifications, the phenomenon of greatest interest is generally the primary characteristic under which a work is classified (and thus determines its shelf placement for physical documents).

---

1 We thus follow Metcalfe, who had distinguished between „specification” (basically, types of a thing), which should be handled by the creation of a hierarchy of subject headings, and „qualification” (everything else: process, aspect form), which should be addressed through linked notation (in Foskett, 1996,127)

2 This is done even when perfect expressivity is impossible, even though it is common in the literature to warn against using only partially expressive notation. I would note that the two-letter codes for US states employed by the US Post office has proved highly useful because it is mostly expressive: states with two-word names are symbolized by the first letters of each word, and most other states by either the first two letters or the first and last. The utility of this system is not erased by the fact that Missouri has to be symbolized by MO (because there are several states starting with MI), though the occasional Missouri-bound letter is likely sent to Michigan by mistake.

3 UDC uses linked notation extensively, and is often criticized for unwieldy notation. Much of the problem stems from very long notation for simple subjects (because UDC builds upon DDC). UDC may also sometimes put more detail in the call number than is required (Foskett 1996, 186)
Kind (of thing). When phenomena are disaggregated into subsidiary phenomena in this classification, this is most often done with respect to „kinds of x...“ When there is a wide range of possible kinds, then linked notation provides a superior classification device: PI4 (CR) is religious organization (PI4 is organization, CR is religion).

Part (of a thing). As with kinds, this facet could be handled by disaggregation (more common in natural science than human science) or linked notation.

Property (of a thing). Y – X means „X of Y“ as in most classifications, and thus captures this facet (but also some relationships among phenomena as well; the meaning of „-“ proves clear in context).

Material (comprising a thing). We would not disaggregate by material but rather use linked notation as with kind of thing above. The precise meaning is again clear in practice because of the existence of recognizable materials within the parentheses.

Process. This facet refers to changes in a phenomenon under study. The BCC uses four easily recalled symbols for changes within a phenomenon: ↑ for increase, ↓ for decrease, ↕ for fluctuations or cycles, and ∩ for stability. Others could be added if necessary.

Operation. Here we capture ways in which a phenomenon is influenced from outside. Most of the classification of relationships (below) will capture this sort of facet.

Product. Here Bliss captures the results of operations on natural entities. Products are for the most part identified in the Basic Concept Classification by the prefix EO9. Such products are otherwise treated as any other phenomenon would be. As noted above, linked notation can be used to identify the material(s) used in construction or the forces that led to the particular product.

By-product. Same as above.

Patient. Bliss refers here to „intermediate goods“ things that are made only to be transformed into something else. These will also be captured within EO9. Intermediate goods can be captured in a causal chain A causes B causes C.

Agent. These are generally individuals, but sometimes tools or institutions. Each of these is classified as phenomena within this classification. Note that we can readily capture different facets in a compact linked notation: <agent> <operation> <product>

Space: Any physical (including political boundaries) or spatial dimension. These are coded as N1 in this classification.

Time: Any chronological or temporal characteristic: These are coded as N2 in this classification.

The Integrative Levels Classification (see www.iskoi.org/ilc) changes the order slightly and adds three new facets:

- Purpose or result. This sort of facet generally describes a relationship and would be captured by a causal relationship: <thing> <relationship> <thing>.
- Pattern (representations, such as a poem about X). This also indicates a compound term and would usually be captured by „of“ or „in“ or „about“
- Modality. This deals with how a phenomenon is perceived. This largely captures elements of the theories and methods employed by scholars. These are clearly signaled by TT and TM. Modality might refer to disciplinary perspective, captured by TF.

Detailed Outline
The entire BCC can be viewed at Szostak (2011c). The classification of phenomena is summarized in Szostak (2011a). The classification of relationships is developed in Szostak (2012).

The guiding principles for classifying phenomena are:
1. The physical world is divided according to the concept of „integrative levels.” At the level of human science, individual-level phenomena are distinguished from societal-level phenomena.

2. Individual-level and societal-level phenomena are organized in a handful of logically distinct categories.

3. Wherever possible, these logical main classes are denoted by the first letter of the class title. Physical libraries are free to choose how they distribute main classes physically. For digital libraries, the mnemonic benefit of this strategy imposes no organizational challenge. Mnemonics are also used for subclasses when possible.

4. More precisely, the classification is grounded in a classification of phenomena developed using a mix of induction and deduction in Szostak (2003b) and reprised in Szostak (2004), and has been expanded significantly both by drawing deductively on other logical classifications (notably the Integrative Levels Classification) and inductively from a variety of sources including ontologies, the DDC, and the works of other information scientists such as Farradane, Perrault, and Khoo.

5. Within each main class, subdivision occurs in terms of „type of” though occasionally „elements of.” It is almost always clear which is the appropriate operative principle (and thus the two need not be distinguished notationally). Subdivision in terms of „elements of” is much more common in natural science, but occurs also in Technology and Science and Health and Population.

6. Subdivision occurs logically. Existing classification systems often abuse hierarchy by including e.g. “applications of x” as hierarchical subdivisions of “x.” In this classification, these inappropriate subclasses are instead captured through compound terminology, generally in the form of causal links.

7. Notationally, main categories and the first level of subdivision are denoted by capital letters, the next level by single-digit numerals, and further levels by lower case letters.

The classification is still a work in progress. In particular, the natural sciences are treated summarily. Nevertheless, it has been found that these schedules are adequate to the task of classifying the bulk of the human science literature at least. The main classes treated so far are:

G Genetic Predisposition
I Individual Differences
E Economy
A Art
S Social Structure
P Politics
T Technology & Science
H Health and Population
C Culture
N Non-human Environment
B Biological Entities [very tentatively]

There is also a class of [adjectival/abverbial] „properties”:
Q Qualities

Following the ILC, it is expected that the following main classes will be developed/borrowed over time: Particles, atoms, molecules (M), celestial, rocks (R).

As noted above, special attention is paid to the classification of methods and theory types (TM and TT within the class „Technology and Science”). There are only about a dozen
methods, broadly defined, that are employed by scholars (or indeed anybody). It is useful, though, to identify particular techniques and approaches within these dozen methods.

Theories are even more challenging to address. As with complex phenomena more generally, theory names carry quite different meanings across disciplines. Not only do quite different theories operate under the same theory name, but similar theories operate under quite different names. Relying exclusively on Cutter numbers to identify theories by name would thus provide limited guidance to scholars curious as to what theories have been applied to a particular causal relationship, or over what range of causal relationships a particular theory has been applied. Theories are thus also identified in terms of a classification of „theory types.” Five key questions are asked of each theory: who is the causal agent, what does that agent do, why does it do that, what is the nature of the causal process that unfolds in the theory, and how generalizable is the theory? A handful of answers are possible for each question. Hundreds of distinct theory types can thus be identified with recourse to only a couple of notational spaces. This approach was developed and applied to grand theories in Szostak (2004, ch. 3) and Szostak (2003a), and was further applied to narrower theories in Gnoli and Szostak (2008). It has been found that this simple classificatory device identifies key differences both across and within theories (while hopefully encouraging scholars to be more explicit about the characteristics of the theory they employ).

As for relationships, guiding principles include:

1. To distinguish relationship concepts from phenomena concepts, mathematical symbols are used rather than letters or numbers notationally. Some of these symbols are commonly employed in other classifications. Italicized letter subscripts are employed for specific types of causal relationship.

2. It is first necessary to classify certain common sorts of non-causal relationships. While many of these must generally be captured in any classification, the emphasis on compound notation in BCC necessitates a longer list. Note that it is not always feasible to search by subsidiary terms in other classifications, nor are these generally standardized across classes. The non-causal relators addressed in BCC are: not/opposite, and/plus, or, of, by/from, for, of type, part of, in, about, compared to, associated with, from the perspective of, of amount, and collection of. Note that various directional comparators can be found under N3, Direction.

3. Since most causal relators have opposites (move/rest, cause/ do not cause), we can reduce the complexity of our classification of relators by allowing the opposite of each (when appropriate) to be captured notationally.

4. Some relators also have inverses: imply is the inverse of infer. It should generally be both possible and desirable to represent inverse meanings in precisely the same way. Since A implied X to B has the same meaning as B inferred X from A, these should be represented by exactly the same notation.

5. Since many relators refer to doing something again (most/all of the „re-“ verbs do so, but there are others), it is also useful to capture all such relators by „do X” combined with the notation for again: N2w.

6. Relators are generally listed in the past participle. It should be stressed, though, that they can generally be used as present participles as well: „A causing B” is similar to „A causes B” for most classificatory purposes. If distinctions needed to be made, this could be done through compounding with suitable temporal indicators.

7. It should also be stressed that the linking terms „to” or „to be” can generally be omitted: „cause to lose” is „cause lose” and „cause to be free” is „cause free.”
An Example

A work describing how “Attitude toward punctuality” affects “economic output” will
within existing systems be classified under one of those phenomena; it may appear under
both in subject indices. The posited causal relationship will not be indicated. In BCC the
work is classified under the entire causal relationship. The notation is nevertheless
compact: CV4f → EO. Note that the notation is also expressive: it tells us that this work
describes influences exerted by a particular cultural value on economic output. If a
particular type of causation were posited in the work, this could be indicated with one or
two italicized letters as a subscript to the arrow. Notation for theory applied or method
applied could be added without the notation becoming unwieldy.

Of course, some sort of notational priority is required for „shelving” within physical
collection of documents. The default order of classification would in fact be the reverse of
that above: Economic output influenced by attitudes toward punctuality; EO ← CV4f. Yet
if it were desired for some reason to shelve such a work with other works on attitudes, the
title and notation could be provided as above. In either case searches by each term in the
compound notation would find the work. That is, in this classification, the notation is
flexible in an important way: X → Y is notationally equivalent to Y ← X but results in
different shelf placement. As well, in the case of complex notation where it would not be
feasible to place the term desired for shelving first in the notation, it is possible to bold a
later term in the notation.

The classifier’s task is simplified mightily in the BCC: they can classify a work under all
relevant causal links, and need only worry about prioritizing a particular phenomenon for
shelving purposes. The user’s task is similarly simplified: the user who knows what causal
link they are interested in can search precisely by that. The user with a less precise query
can be guided to a list of links that they might investigate. Searches that uncover all
relevant works for both types of user are more likely than at present to yield advances in the
discovery of „undiscovered public knowledge” or „serendipity.”

Benefits

A few of the key benefits that flow from this classification can be briefly noted:

1. Since most works can be classified as links between phenomena, we are able to
achieve very precise classifications with limited and expressive notation.
2. Users are thus better able to find precisely what they want, whether they wish to
search in one discipline or across all. Works need not but can be coded by discipline.
3. By distinguishing different sorts of relationship (especially causation/influence), we
enable searches by verb-like terms as well.
4. While other classification systems provide specific instructions in multiple places for
coding by time or place or people, this system has a universal coding for such
elements. This renders both classification and searching easier.
5. At times, there is some ambiguity as to how to code a complex causal relationship
(e.g. terrorism). [Thus, there is sometimes hidden dual coding, so that the term can be
found using multiple symbolic searches.] Users may often find it useful to enter the
classification by entering search terms in the usual way. Nevertheless, the
classification can still alert users to closely related linkages.
6. The system is ideally suited to Boolean searches using symbols [though users can
achieve similar results by searching by terms; the computer can translate words to
symbols]. A user searching by C plus → plus P will uncover all of the ways that any
aspect of culture has any sort of influence on any aspect of politics. They can then
easily narrow their search.
7. Note that the use of linked notation serves to place works within multiple hierarchies (and of relations as well as things)
8. Note likewise that a document can be given a very detailed classification, but also a much simpler shelf-mark that reflects its main phenomenon or causal linkage.
9. It should be possible to translate all search or entry terms employed in other classifications into basic concepts. Note that in addition we create the possibility of (fairly) automatically coding for new works and for existing works that are at present poorly classified.

References
Szostak, Rick. 2011b. “Translation Table: DDC to BCC. Website. http://www.economics.ualberta.ca/en/FacultyAndStaff/~media/economics/FacultyAndStaff/Szostak/Szostak-Dewey-Conversion-Table.pdf
Abstract
Aiming at contributing to the epistemological characterization of the area of knowledge organization, our goal is to analyze the KO journal, since its creation in 1993, as a knowledge domain, from a nuclear community of the most productive and greater impact authors, analyzing the dialogue among citing authors and cited ones, and also the co-citations established by the citing authors. We worked with a corpus of 310 articles published between 1993 and 2011 produced by a total of 360 authors. The relatively more productive authors, a group geographically concentrated in Europe (37%), North America (44%) and Asia (19%), is clearly explained by the historical European origin of the ISKO and by an increasing North American presence along the years. Of the 33 most cited authors, 22 were co-cited in at least 6 works, which suggests that they are the theoretical referential nucleus of the area, in the studied journal. Finally, we observe that the area reveals theme cohesion and coherence in its production, enabling us to clearly visualize its theoretical referential nucleus and to confirm the role performed by the KO magazine as a catalyzing agent of international theoretical construction in the area.

Introduction
Placed among Cognitive Science, Information Science, Communication Science, Math and Computer Science, Knowledge Organization can be characterized by its social and scientific nature, revealing itself as an inter and trans disciplinary field (Garcia Marco, 1995, 1997; Miranda, 1999) and evidencing an “integration platform of the documentary sciences” (Esteban Navarro, 1995, p.66).

Therefore, Knowledge Organization aims at investigating how knowledge can be understood, organized and represented, so that it is available and can be retrieved by the greatest number of people. To do so, it presents a tridimensional nature, centering itself on principles, methods and tools to manage human knowledge from its representation, organization and documentary communication (Esteban Navarro, 1995; Dahlberg, 2006).

In this sense, the International Society for Knowledge Organization – ISKO has a special contribution since it is considered a space for scientific discussions about knowledge organization as well as by the study of different methods and approaches in order to facilitate access to knowledge by the society. To do so, ISKO has Knowledge Organization journal – KO as its main socialization vehicle for scientific production.

Willing to contribute to the epistemological characterization of the knowledge organization area, our goal is to characterize the KO journal, since its creation in 1993, as a knowledge domain, from a nuclear community of the most productive authors. More specifically, from these authors and their citations, we intend to visualize the researchers of greater insertion and impact, analyzing the dialogue between citing authors and cited ones, as well as the co-citations established by the citing authors.

Domain analysis, the object of this work, has been discussed in the ISKO scientific environment on several occasions. Smiraglia (2011) has presented a critical synthesis of the previous works related to knowledge organization as a domain, in different periods and bibliographic sources. In subject approach, McIlwaine (2003), Saumure & Shiri (2008) and Ibekwe-Sanjuan & San Juan (2010) have studied output in different periods involving journals and proceedings. Regarding the ISKO international congresses, Smiraglia (2006, 2008, 2011c) analyzes the international congresses of Vienna, Montreal and Rome,
respectively. Within regions, we have an analysis of the North American context from the NASKO congresses (Smiraglia, 2007, 2009), of the Spanish context (López-Huertas e Jimenéz-Contreras, 2004) and of the Latin American context (Smiraglia, 2011b).

This current work assumes the whole collection of KO journal as a knowledge domain in itself and presents it as a differential, in that it reveals a crystallized international space for scientific publishing with the seal of ISKO, thus representing the research tendencies in the area. Thus, the specific analysis of the theoretical referents coming from the predominant citations in such knowledge domain intends to identify, in a diachronic study, to which extent this scientific community bases itself on and interacts in the construction of this specialized knowledge.

**Knowledge Organization Journal: A Domain Analysis**

Although authors like Dahlberg (1993, p.214) have pointed out the individual concept of knowledge as “subjective certainty or objective conclusion of the existence of a fact or state of a case, not being transferrable and only acquired by means of reflection”, knowledge organization, within Information Science, focuses mainly on socialized, recorded and published knowledge, whose organization and representation will be developed in a way that, new knowledge may be generated from it (Miranda, 1999; Barité, 2001; Guimarães, 2001). Such aspect reveals a helicoidal concept – not cyclical as it was used to say until some years ago – of knowledge organization. Thus, knowledge acts as a product, need and social dynamo, which occurs from the information, and while being socialized, turns into it again (Dahlberg, 2006; Guimarães, 2008).

Hjørland (2003) warns that knowledge organization, within Information Science, takes a deeper and stricter focus, having the semantic relations between concepts as its basic unit, and not the concepts themselves.

Intending to congregate knowledge organization researchers worldwide, ISKO has tried to improve, since its creation in 1989, philosophical, psychological and semantic approaches to knowledge organization, also acting as a bond among national and international societies and institutions, about issues related to conceptual nature of knowledge organization and processing. In this context, Knowledge Organization journal, created in 1993 (continuing International Classification, created in 1974) can be considered the main scientific publishing vehicle of the researches made in the area.

ISKO, in the last decade, has notably revealed certain concern by the scientific community about the epistemological construction of the area, that led to the publication of a KO special number, in 2008, about this theme, having the contribution of Ingetraut Dahlber, Claudio GnoLi, Rebeca Green, Birger Hjørland, Maria José Lopéz-Huertas, Ia McIlwaine, Joan Mitchell, Joseph Tennis and Marcia Zeng, adding to other studies of Smiraglia (2011b). That confirms previous observations of Hjørland (1994, 2003), Kiel (1994), Jaenecke (1994), Barité (2001) and Smiraglia (2002) as for the need to dedicate greater efforts to mapping the area, its structure and limits, as it’s thought, many times, dependent on the progress of information and communication technologies rather than on its own theoretical research.

To do so, the theoretical and methodological support of Domain Analysis is necessary to identify, visualize and better understand the theoretical milestone universe which surrounds the area.

Hjørland (2002) highlights 11 approaches by which a domain can be analyzed. Bibliometric studies are one of them, and are specially focused on citation and co-citation as means to visualize the domain behavior in terms of its authors and the knowledge produced and socialized by them. Thus, the group of references of the scientific works, studied by means of citation and co-citation analysis of frequency, contributes to domain visualization as the reflection of a discursive community.
A citation is taken as an objective and clear indicator of scientific communication, which allows the identification of scientists and their publications, in order to evidence the researchers with greater impact in the area. It is also important to pointing out to their paradigms, methodological procedures, as well as to the “old school” researchers who build new knowledge in the area. This way, citation analysis contributes to the understanding of a scientific community as it identifies the researchers with greater impact in the area and gives visibility to their concepts, methods, objects and to the theoretical references that support such community (Oliveira, Gracio e Silva, 2010).

Therefore, “citation analysis maps the scientific communication” (Vanz, Caregnato, 2003, p.248) and provides indicators that contribute to the construction of networks, making the dialogue among researchers clearer. This way, it constitutes a special tool so that, from quali-quantitative analyses, we may better understand the epistemological universe of a certain domain.

The study of co-citations, which comes from the citation analysis, deals with the frequency with which two authors or documents are cited within a group in the scientific production of an area, producing valid representations of the intellectual structure of a scientific domain (Miguel, Moya Anegon and Herrero Solana, 2008). Still according to the referred authors, the main premise of the co-citation analysis is that if two or more documents, authors or journals are cited together in a third following work, there is, at least in the citing author’s perspective, a similarity of subjects among the cited ones. We point out that the greater the frequency of co-citations is, the closer the relationship among them. Not only the similarity, but also the opposition of ideas is detected by means of co-citation. According to Gmüer (2003), the co-citation frequency between two authors also determines how the knowledge structure of an area is perceived by its researchers.

The importance of citation and co-citation analysis as an approach to characterize a domain is more clearly evidenced in that the repeated citation of certain works, by a given scientific community, reveals an agreement in this community about what to consider relevant a priori, characterizing a trait of that certain discursive community, once the structure of a certain scientific field is built from patterns of co-recognition, evidenced by means of co-citations, which establish meaningful associations (Small,2004).

Methodological Procedures

To raise the data, we firstly searched in the Web of Science database, the group of articles published in KO journal. To this group of documents raised, we applied the filter “Articles” in “Document Types” acquiring a group of 310 articles published between 1993 and 2011 and produced by a total of 360 authors.

From this group, we identified the most productive authors, a total of 16 researchers (4.4% of the total of authors) who produced a total of 57 articles (18.3% of the published articles)\(^1\). The adopted criterion for the selection of such was the publication of at least 3 articles in the period, once the authors with two publications or fewer comprised a large, spread and less meaningful group.

To raise the references for the 57 articles, we used the digital collections of Ergon Verlag, available for ISKO members’ access and also the references of the articles in the Scopus base, since this base brings the references with all the authors, not just the first author as in the Web of Science base. We obtained a group of 1755 references and, after eliminating the references with institutional authorship, we obtained 1635 citations, referring to 1007 authors, from which 758 (75,3%) received only one citation.

As a criterion to include a name in the roll of most cited authors, we considered the ones with at least 6 citations (the same as being cited in at least 10% of the works in the corpus), which gave us a group of 37 most cited researchers. From these, we built a matrix of

---

\(^1\) Six articles were excluded from the work corpus because they didn’t bring bibliographic references.
16x37, related to the frequency of citations of the most productive citing authors (16) and the most cited ones (37).

From this matrix, we generated a squared and symmetrical matrix 22x22 with co-cited researchers in at least 6 works, corresponding to the co-citation in at least 10% of the articles. We built the co-citation network by means of the Ucinet software in order to map and visualize the communication network established among researchers in the citing works. After that, we calculated the centrality degree, which enables to analyze the individual position of each author, as well as the most meaningful and articulated authors in their group’s network.

**Data Presentation and Analysis**

Regarding the 16 most productive authors, we have, in decreasing order of production: Hjørland (06); Dahlberg (05); Fugmann, Gnoli, Mai, Riggs, Tennis, Beghtol (04 each) and Albrechtsen, Chaudry, Green, McIlwaine, Olson, Pathak, Satija, Zins (03 each). This group is geographically represented by a scientific production coming from Europe (37%), North America (44%) and Asia (19%), what can be historically explained by the European origin of the ISKO and by an increasing North American presence along the years. This group of most productive authors signals a group that has contributed to the construction of knowledge in the area hereby studied in a more meaningful way.

**Table 1: Most productive researchers’ works in which the most cited ones are shown**

<table>
<thead>
<tr>
<th>Most productive researchers</th>
<th>Most cited authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hjørland</td>
<td>6</td>
</tr>
<tr>
<td>Dahlberg</td>
<td>5</td>
</tr>
<tr>
<td>Fugmann</td>
<td>4</td>
</tr>
<tr>
<td>Gnoli</td>
<td>4</td>
</tr>
<tr>
<td>Mai</td>
<td>3</td>
</tr>
<tr>
<td>Riggs</td>
<td>2</td>
</tr>
<tr>
<td>Tennis</td>
<td>2</td>
</tr>
<tr>
<td>Beghtol</td>
<td>1</td>
</tr>
<tr>
<td>Pathak</td>
<td>1</td>
</tr>
<tr>
<td>Albrechtsen</td>
<td>0</td>
</tr>
<tr>
<td>McIlwaine</td>
<td>0</td>
</tr>
<tr>
<td>Olson</td>
<td>0</td>
</tr>
<tr>
<td>Green</td>
<td>0</td>
</tr>
<tr>
<td>Zins</td>
<td>0</td>
</tr>
<tr>
<td>Chaudry</td>
<td>0</td>
</tr>
<tr>
<td>Satija</td>
<td>0</td>
</tr>
<tr>
<td>Tot. citations</td>
<td>5</td>
</tr>
</tbody>
</table>

**Note:**

- Tot. citations: Total citations of the most productive authors.
- Hjørland: Total citations for Hjørland.
- Dahlberg: Total citations for Dahlberg.
- Fugmann: Total citations for Fugmann.
- Gnoli: Total citations for Gnoli.
- Mai: Total citations for Mai.
- Riggs: Total citations for Riggs.
- Tennis: Total citations for Tennis.
- Beghtol: Total citations for Beghtol.

**Table 1:** 31 rows, each row representing a researcher's name and the number of citations they have. The table shows the total citations for each researcher, with the researchers being ranked in descending order of citations.
Most productive researchers

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pathak</td>
<td>1   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>1   2   0   1</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Albrechtsen</td>
<td>0   0   0   0</td>
<td>2   1   1   1</td>
<td>0   0   0   0</td>
<td>1   1   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>McIlwaine</td>
<td>0   0   0   1</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   1   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Olson</td>
<td>0   0   1   1</td>
<td>0   2   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>1   0   0   0</td>
<td>0   1   0   0</td>
<td>0   0   1   0</td>
<td>0   0   0   0</td>
<td>1   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Green</td>
<td>1   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>1   0   0   0</td>
<td>0   0   0   0</td>
<td>1   0   0   0</td>
<td>0   1   0   0</td>
<td>0   0   0   0</td>
<td>0   1   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zins</td>
<td>0   0   1   0</td>
<td>1   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>1   0   0   0</td>
<td>1   0   0   0</td>
<td>1   0   0   1</td>
<td>1   0   0   0</td>
<td>1   1   0   2</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chaudhry</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   3   0   1</td>
<td>1   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Satija</td>
<td>0   0   0   1</td>
<td>0   1   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td>0   0   0   0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tot.citations *</td>
<td>7   1   0   9</td>
<td>9   7   7   9</td>
<td>8   21   7   6</td>
<td>13   6   12   6</td>
<td>8   6   8   8</td>
<td>9   6   6   6</td>
<td>6   6   6   6</td>
<td>6   6   6   6</td>
<td>6   6   6   6</td>
<td>6   6   6   6</td>
<td>6   6   6   6</td>
<td>6   6   6   6</td>
<td>6   6   6   6</td>
<td>6   6   6   6</td>
<td>6   6   6   6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The total citation refers to the number of works where the author was quoted in the universe of 57 articles, disregarding the co-authorships

This group of theoretical referents is confirmed by the fact that, of the 16 most productive authors, 12 (75%) are equally pictured as the most cited researchers.

Among the most productive researchers, Mai, Beghtol, Hjørland, Gnoli and Tennis stand out as the ones who establish greater dialogue with the group of 33 most cited authors, since they cite most of the authors, between 50% and 80%.

Hjørland stands out as the most productive and the most cited author, being mentioned in 40% of the analyzed articles, that shows his recognition and visibility among the highlighted researchers in the journal studied. Ranganathan is cited in 37% of the articles and Dahlberg in 30% of the articles. The other researchers were cited in between 26% and 11% of the articles, in decreasing order. We highlight the presence of the classical authors of the area, such as Bliss, Langridge, Vickery, Foskett and Dewey.

Hjørland and Ranganathan were cited by 56% of the most productive authors. Following them are Albrechtsen and Vickery, cited by 44% and Dahlberg, Beghtol and McIlwaine, cited by 38% of the researchers. We understand that these most cited authors constitute the focus of greater impact and prominence.

In all the cases, self-citations were not taken into account. Although present, they were not significant compared to the total of citations received, for most authors, who are the most productive and most cited ones altogether, not causing any interference in the analysis of insertion and impact of most cited authors.

Figure 1 presents the co-citation network among authors who were co-cited at least 6 times. So, 22 of the 33 most cited authors were co-cited in at least 6 works, what reveals that 67% of the most cited authors are part of this network suggesting that they constitute the core of theoretical referents in the area, in the journal studied.

We highlight Hjørland with the greatest centrality (71%), being co-cited with 15 of the authors present in the network. Ranganathan, with 39%, is co-cited with 8 authors and Albrechtsen and Beghtol, with 33%, is co-cited with 7 authors.
Conclusions

It is possible to conclude that there is a strong nucleus of 16 theoretical referents, whose origin reveals a balance between Europe and North America, historically considered as privileged spaces of knowledge organization research.

The presence of authors like Hjörland, Ranganathan and Dahlberg reveals a predominant theoretical approach, more tuned to the search for epistemological foundations of knowledge organization that confirms a tendency foreseen by Gnoli, 2008 and Lopéz-Huertas, 2008, among others. The referred results reveal a dialogue with the previous researches cited, in that authors, such as Hjörland, Ranganathan, Dahlberg, Begthol, Gnoli and Vickery are equally present in the studies carried out by Smiraglia, previously cited.

Finally, we observe that the knowledge organization area, taking into account its scientific production, reveals balance and diversity concerning the geographical origin, as well as subject cohesion and coherence, allowing us to clearly visualize the nucleus of its theoretical referents and confirming the effective role performed by the KO journal as a catalyzing agent of the theoretical construction in the area worldwide.

References


José Augusto Chaves Guimarães - UNESP, Brazil
Joseph T. Tennis - University of Washington, USA

Constant Pioneers:
The Citation Frontiers of Indexing Theory in the ISKO International Proceedings

Abstract
Presents a citation analysis of indexing research in the ISKO Proceedings. Understanding that there are different traditions of research into indexing, we look for evidence of this in the citing and cited authors. Three areas of cited and citing authors surface, after applying Price’s elitism analysis, each roughly corresponding to geographic distributions.

Introduction and Background
Reflecting on past research in indexing is beneficial for both the theory and the practice of indexing for a number of reasons. First, we want a greater understanding of the contours of indexing theory. Second, we want to know what has been said so we can think about it and then act with knowledge. Third, we want to be able to evaluate indexing theory, and finally we want to set the stage to develop theory further. The process of examining indexing theory this way is called metatheory (Ritzer, 1991).

In order to reflect on past research on indexing, we gathered authors researching indexing and publishing in the ISKO context from 1990-Present, and extracted citations in these papers. We wanted to see what the landscape of indexing theory looks like over time through the literature cited.

It is important that we better understand the literature of indexing because we can see that there are at least three traditions: (1) subject cataloguing, (2) indexing, (3) and analyse documentaire (Guimarães, 2008). These three traditions draw on separate literatures as well as national traditions. So it is important for indexing theory in the international ISKO context to study the landscape of the literature.

Subject cataloguing had its origins in North America, specially from Cutter’s principles of alphabetical cataloguing, the tradition of LC subject headings and the thoughts of the School of Chicago, whose emphasis is mostly focused on the catalogue - a product of information treatment in libraries. Such conception is defined by Fiuza (1985, p. 257) by “... representing, in library catalogues, the subjects contained in the collection”. We can count Cutter, Kaiser, and Coates as core thinkers in this tradition. Currently, authors such as Hope Olson and Sanford Berman, among others, have dedicated to the aspects of subject cataloguing. This approach has found in periodicals as Cataloguing and classification quarterly, a special vehicle, considering it being mainly related to questions of organization of bibliographical registers and the bibliographical control in general.

The second approach - Indexing – comes from a British tradition where indexes, while products of the information subject treatment, are pinned down from the usage of indexing languages – specially thesauri - in specialized information centers or libraries, with a deep concern on the theoretical nature about the construction of such languages, is much influenced by the work of Classification Research Group. The contributions of Foskett, Austin, Farradane, Metcalfe, Aitchison, Gilchrist and Lancaster stand out. The journal The Indexer can be considered a special vehicle for spreading this conception, considering that it predominantly devoted to issues concerning history, organizations, systems, standards, methods, practices and technology indexing.
Another dimension, called *analyse documentaire*, comes from the French tradition, whose focus is centered on the subject treatment process itself, by means of the explanation of the procedures involved in identifying and selecting concepts. Those studies had a strong emphasis from the late 60's, with deep interface with Logics, Linguistics and Terminology, from the pioneering work of Gardin (1966a, 1966b, 1967, 1970, 1973, 1974, 1981) and of Coyaud (1966). Journals like Documentaliste and Journal of Documentation, can be considered as representative of such theoretical approach, especially by considering their tradition more focused on the documentation matters as well as on the methodological questions involving content analysis. In the *analyse documentaire* approach it is possible to observe how important the theoretical questions surrounding the document are, by revealing the influence of both the Diplomatics discussions (Tessier, Bautier, etc.) from the École de Chartes from the second half of the 19th Century until the beginning of the 20th Century and the Documentation concepts (Otlet, Briet) in the first half to the 20th Century.

Now we can ask an empirical question: do we see these three dimensions of indexing in the ISKO proceedings?

**Methodology**

Drawing on all of the proceedings of ISKO International Conferences, we searched for papers that had the word *indexing* in the title or abstract. We then recorded the author’s name and extracted the citations for each of these papers (excluding self-citations and standards), and from this data we created citation networks for each ISKO proceedings volume and also cumulatively for all the proceedings. We found a total of 784 cited authors.

For the creation of the citation networks we used the software PAJEK which allows for the construction of graphic network displays for the citations. From this data, we can describe the landscape and contours of indexing theory against the backdrop of these national and intellectual traditions of subject cataloguing, indexing, and analyse documentaire, and specifically talk about the ISKO environment.

**Analysis and Findings**

The most highly cited authors, those in the 98th percentile are (in rank order): Hjørland, B., Lancaster, F.W., Vickery, B. C., Aitchison, J., Beghtol, C., Larson, R. R., Ranganathan, S. R., Soergel, D., Reisthuis, G. J. A., Cochrane, P. A., Croft, W. B., Saracevic, T., Gilchrist, A., Svenonius, E., Chan, L. M., Hearst, M. We see a large representation of the British and North American thinkers, and a relative lack of the French tradition. In fact, only five citations total for Gardin, the only representative of the French school present in these ISKO proceedings.

We then applied Price’s Theory of Elitism (Price, 1963) to the 784 cited authors. This theory supposes that the elite of a certain citing environment is composed by the square root of the number of cited authors:

\[
\text{Elite } X = \sqrt{X} \quad \text{(Braga, 1974)}.
\]

In this sense, and considering the original group of cited authors, it was possible to arrive at a group composed of 28 cited authors: Albrechtsen, Atchison, Bates, Beghtol, Broughton, Chan, Coates, Cochrane, Croft, Gilchrist, Hearst, Hjørland, Lancaster, Larson, Markey, McIlwaine, Mitchell, Olson, Politt, Ranganathan, Reisthuis, Salton, Saracevic, Soergel, Svenonius, Tobias, Vickery, Vizine-Goetz. The original group of 127 citing authors was transformed in a group of 87 citing authors who have cited the elite authors.

Using this group of elite cited authors and the corresponding citing authors, we built a citation network using PAJEK. (See Figure 1).
In the mentioned group it is possible to observe that 3 authors - Albrechtsen, Cochrane and Politt belong at the same time to the categories of most productive (citing authors) and most cited authors. The remaining 25 most cited elite authors surround them, while the citing authors (N=84) surround them.

It is also possible to observe a high level of relationship among the cited authors, revealing a high density level in the network. It allows us to conclude that subject indexing presents a strong cohesion in terms of theoretical referents. That is, very few cited authors stand alone.

Only one of the 28 most cited authors – Larson - does not establish relationships with the others, and so constitutes a separate network with his citing authors (Bartolo, Kin, Buckland, Norgard, Chen and Trimble).

The universe of most cited authors (Price’s elite) reveals different theoretical approaches. In this sense, authors like Aitchison, Broughton, Coates, Gilchrist, Lancaster, McIlwaine, and Vickery represent a traditional “English approach on indexing” (Guimarães, 2009), with a strong relationship and influence of the Classification Research Group – C.R.G. On the other side, authors like Hjorland, Soergel, and Riesthuis, are more connected to the theoretical and conceptual basis of knowledge organization - KO, representing the ethos and agenda of investigating the fundamental basis of ISKO creation. In this environment, Ranganathan acts as a connection between the mentioned approaches, since he has deeply influenced the English tradition of C.R.G. as well as Dahlberg’s and Soergel’s conceptions of KO.
The North-American tradition is also significant in the analyzed network. The authors Bates, Beghtol, Mitchell, Olson, Saracevic, and Svenonius establish a strong relationship between the theories of subject cataloguing and information retrieval and confirming the presence of North American authors in ISKO literature as observed by Smiraglia (2009).

The one outlier is Larson, connected with information retrieval, but not in dialogic relationship with any other literature in ISKO. And as noted above, the French tradition does not figure in the proceedings of ISKO.

Conclusion

This small contribution to the metatheory of indexing comports with similar studies of different indexing theories (Tennis, 2005); that is to say, there is a great diversity of thought indexing. We can add to what was described above, there are more than three traditions: a) subject cataloguing, based on the North American tradition of School of Chicago and mostly focused on building catalogues for public and school libraries; b) indexing, based on the British tradition of the Society of Indexers as well as of the Classification Research Group, mostly focused in building indexing languages for representing knowledge in special libraries and documentation centers, and c) analyse documentaire, based on the French tradition and mostly focused in explaining the logic and linguistic procedures involved when representing knowledge in special libraries and documentation centers (Guimarães, 2008). And two of these are present in ISKO. The other seems conspicuously absent. The next question then, is why, especially given the prominence of Otlet and Briet in our current conceptions of the field. We can add to this list the North Americans concerned with information retrieval (i.e., Larson), and those that combine subject cataloguing and information retrieval (i.e., Bates, Beghtol, Mitchell, Olson, Saracevic, and Svenonious).

What is clear from this part of the study is that indexing theorists are constant pioneers, constantly charting new territory by incorporating new literature into the field of indexing, and maneuvering in particular traditions. On the other hand, it is possible to observe that although those traditions are focused in different points, they establish dialogical relationships in the KO environment – maintaining their position in their citation field, but exploring new frontiers as the research advances into unknown territory.

References


The Relationship between Authors and Main Thematic Categories in the Field of Knowledge Organization: A Bibliometric Approach

Abstract
This is a study about the relationships between authors and the main thematic categories in the papers published in the last five International ISKO Conferences, held between 2002 and 2010. The aim is to map the domain as ISKO conferences are considered the most representative forum in the field. The published papers are considered to indicate the relationships between authors and themes. The Classification Scheme for Knowledge Organization (CSKOL) was used to categorize the papers. The theoretical and methodological foundations of the study can be found in the concept of domain analysis proposed by Hjørland. The analysis of the papers (n=146) led to the identification of the most productive authors, the networks representing the relationships between the authors as also the categories that constitute the primary areas of research.

Introduction
Considering knowledge organization (KO) as a domain, for which the continuous construction of a consistent epistemological statute is necessary, this research seeks to identify the main areas of thought and research in the domain. This is done based on the papers presented in the international conferences of the ISKO so as to establish relationships between researchers and the themes. The research also aims at identifying the more productive authors in KO based on papers in ISKO international conferences, as well as examining the themes these researchers focus on. It also aims at identifying the themes that were of greater interest in the last five editions of the event, held between 2002 and 2010.

The choice of papers in ISKO International conference volumes for domain analysis was based on the assumption that these are representative of the research in the domain. As such an examination of the papers should contribute to an understanding of the real and present composition of the domain.

Knowledge Organization: A Bibliometric Analysis
The domain of KO has for its foundation the philosophical framework of Dahlberg’s concept theory (1993), the author considered to be the pioneer in the field of studies related to KO. She is also mainly responsible for forming ISKO. The construction of classificatory and thesauri systems have been structured based on this conception, which demonstrates that the concepts theory is widely accepted in the area. It can therefore be said that Dahlberg’s theory plays an important role in the process of cognitive institutionalization of the discipline.

Based on the concepts theory, Dahlberg (1993) has developed Classification Scheme for Knowledge Organization Literature (CSKOL) at ISKO, which groups the main concepts under ten primary categories: C0 Form Divisions; C1 Theoretical Foundations and General Problems; C2 Classification Systems and Thesauri (CS&T): Structure and Construction; C3 Methodology of Classing and Indexing; C4 On Universal Classification Systems and
In order to better comprehend the domain of KO and the performance of its researchers in the thematic categories defined in CSKOL, the use of the theoretical-methodological foundation of domain analysis is necessary. To Hjørland and Albrechsten (Hjørland, 2002, 2003; Hjørland & Albrechsten, 1995) (1995), a scientific domain is understood as the reflection of a discursive community and its role in science. In this perspective, Hjørland (2002) says that there are eleven approaches to domain analysis, including the bibliometric approaches, which constitute consistent approaches in analyzing and characterizing a scientific domain. Besides this, he also highlights that the joint use of more than one of these approaches enriches the analysis and comprehension of a domain.

In the last few years, the application of social networks analysis methodology in these studies has quickly been consolidated to visualize and highlight the existing connections between researchers, fields, themes, institutions and geographical areas in a knowledge domain. This application, associated to other approaches, such as the epistemological, critical and historical studies, enables a larger comprehension and analysis of a scientific community. Specifically in this research, the analysis of social networks favors the representation of KO, depicting the content produced in the last five international conferences of ISKO. This entity is effectively a social institution which furthers the development of research in KO, as it constitutes a privileged forum for researchers to progress together and validate the theoretical fundaments of the area.

As affirmed by Wasserman and Faust (1994, p. 9), “the term ‘social network’ refers to the set of actors and the connections between them”. Network analysis has the modeling of the relationships between actors as its goal, aiming to portray, describe and represent the structure of a group. According to Otte and Rousseau (2002), the analysis of social networks is a method which enables the mathematical visualization of the social relationships of a certain field, and it is because of this, that it has been applied in several areas of knowledge. According to the authors, it is possible to verify how the structures of each domain may influence the behavior of researchers through social networks analysis and, effectively describe how the thematic tendencies of citations and co-citations of a certain field are developed. This study, in particular, uses the relationship network between researchers and themes in KO, as well as thematic co-occurrences.

Using the social network analysis, Oliveira, Grácio and Silva (2010) investigated the social relationships in the domain of organization and knowledge representation in the Brazilian context. The authors detected a convergence between the contents studied by the authors, as well as a large similarity and/or complementarity of contents which allowed the visualization of the most hegemonic research lines in such domain.

Considering Hjørland’s (2003) warning regarding the lack of theoretical mapping of the discipline, which was reinforced by Guimarães (2008), this study aims to contribute to the outlining of the most representative themes in KO at an international level.

**Methodological Procedures**

The methodological procedure was the listing of a set of 311 papers presented by 524 authors, published in the last five ISKO international conference volumes that took place in

The register of all published papers in the five conference was done in an Excel spreadsheet which contained information about the authors, titles and the corresponding conference event. Only the authors who produced three or more articles during the period of analysis were considered, resulting in a total of 43 authors, responsible for 146 papers. From this selection, 13 papers involving more than one author were identified - 12 papers with two authors and one paper with three authors.

The most productive authors were related to the CSKOL categories, based on the themes of the presented work. Wherever necessary a paper was assigned to more than one category so as to include all the thematic facets of each work. The number of classification codes assigned to a paper varied between one and five. A relationship matrix between these two variables was generated by entering the data into a 43x9 matrix (authors x categories). This structure enabled the data to be exported for further processing using the Ucinet software in order to create the relationship graphs between authors and CSKOL categories (Figure 1). In order to verify the correlation between the nine CSKOL categories throughout the 146 works, a 9x9 table was constructed indicating frequencies of co-occurrence (of categories).

Results Analysis

Table 1 presents the set of 43 researchers with a higher level of participation at the event, signified by at least three papers.

The first four have six papers each and three of them are from North-American universities. Most authors (22) from Table 1 come from American and Canadian universities, followed by seventeen authors affiliated to European universities, three to Latin-American and one to Asian, the last being among the four most productive ones. Such results show that, quantitatively speaking, North-American institutions of teaching and research in KO have a large presence in the ISKO international conferences.

<table>
<thead>
<tr>
<th>Researcher</th>
<th>Total</th>
<th>Researcher</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green, R. (E.U.A.)</td>
<td>6</td>
<td>Beghtol, C. (Canada)</td>
<td>3</td>
</tr>
<tr>
<td>Jacob, E. (E.U.A.)</td>
<td>6</td>
<td>Coleman, A. (E.U.A.)</td>
<td>3</td>
</tr>
<tr>
<td>Khoo, C. (Singapore)</td>
<td>6</td>
<td>Dousa, T. (E.U.A.)</td>
<td>3</td>
</tr>
<tr>
<td>Smiraglia, R. (E.U.A.)</td>
<td>6</td>
<td>Frâncu, V. (Romenia)</td>
<td>3</td>
</tr>
<tr>
<td>La Barre, K. (E.U.A.)</td>
<td>5</td>
<td>Furner, J. (E.U.A.)</td>
<td>3</td>
</tr>
<tr>
<td>Lee, H.-L. (E.U.A.)</td>
<td>5</td>
<td>Herrera-Viedma, E. (Spain)</td>
<td>3</td>
</tr>
<tr>
<td>López-Huertas, M. (Spain)</td>
<td>5</td>
<td>Hjorland, B. (Denmark)</td>
<td>3</td>
</tr>
<tr>
<td>Tennis, J. (E.U.A.)</td>
<td>5</td>
<td>Horwath, L. (Canada)</td>
<td>3</td>
</tr>
<tr>
<td>Broughton, V. (United Kingdom)</td>
<td>4</td>
<td>Hudon, M. (Canada)</td>
<td>3</td>
</tr>
<tr>
<td>Gnoli, C. (Italy)</td>
<td>4</td>
<td>Ibekwe-SanJuan, F. (E.U.A.)</td>
<td>3</td>
</tr>
<tr>
<td>Guimarães, J. (Brazil)</td>
<td>4</td>
<td>Kobashi, N. (Brazil)</td>
<td>3</td>
</tr>
<tr>
<td>Hadju Barát, A. (Hungary)</td>
<td>4</td>
<td>Loehrlein, A. (E.U.A.)</td>
<td>3</td>
</tr>
<tr>
<td>Melllweine, I. (United Kingdom)</td>
<td>4</td>
<td>Ohly, H. P. (Germany)</td>
<td>3</td>
</tr>
<tr>
<td>Moya-Anegón, F. (Spain)</td>
<td>4</td>
<td>Ou, S. (United Kingdom)</td>
<td>3</td>
</tr>
<tr>
<td>Naumis Peña, C. (Mexico)</td>
<td>4</td>
<td>Priss, U. (Germany)</td>
<td>3</td>
</tr>
<tr>
<td>Olson, H. (E.U.A.)</td>
<td>4</td>
<td>Qin, J. (E.U.A.)</td>
<td>3</td>
</tr>
</tbody>
</table>

1 Exception made to the papers published in the annals of the 2010 edition, held in Rome, as they had already been classified by the event organizers.
In Table 2, the frequencies relating to papers distributed within the nine categories can be observed, highlighting C1. Considering the fundamental theoretical character of this category, this strong presence is explained by the necessity felt by researchers to reassess theoretical aspects referring to the domain. This need has the intent of verifying if its theoretical fundamentals are still able to support the empirical reality, the concrete transformations of which have become more evident making researchers reflect on these. The themes approached, such as the concept of order, conceptology, mathematics, systems theory, psychology, sciences in general, all apply to KO. This, as well as the objects and problems of research related to classing and the history of the domain itself are all present in the analyzed research.

Categories 7, 9 and 2, relating to the application of knowledge and concrete questions about KO, are also stressed. In C7, themes relating to the representation of knowledge through language and terminology are approached, whereas C9 presents issues referring to social context, highlighting those associated to people’s influence, institutions and, effectively, their ethical and ideological attitude in the practice of KO. C2 is related to the structures, construction elements and methods, maintenance, update and validation of classification systems and thesauri. The high incidence of these categories shows the growing concern with the development and structuring of knowledge organization systems in the semantic web sphere and its impact in social contexts, which is a recurrent theme in the research presented at ISKO conferences.

**Table 2: Frequency of categories occurrence in works**

<table>
<thead>
<tr>
<th>Category</th>
<th>Total</th>
<th>%*</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1: Theoretical Foundations and General Problems</td>
<td>58</td>
<td>39,7</td>
</tr>
<tr>
<td>C2: Classification Systems and Thesauri (SC&amp;T): Structure and Construction</td>
<td>44</td>
<td>30,1</td>
</tr>
<tr>
<td>C3: Methodology of Classing and Indexing</td>
<td>16</td>
<td>11,0</td>
</tr>
<tr>
<td>C4: On Universal Classification Systems and Thesauri</td>
<td>18</td>
<td>12,3</td>
</tr>
<tr>
<td>C5: On Special Objects Classification (Taxonomies)</td>
<td>3</td>
<td>2,1</td>
</tr>
<tr>
<td>C6: On Special Subjects Classifications and Thesauri</td>
<td>14</td>
<td>9,6</td>
</tr>
<tr>
<td>C7: Knowledge Representation by Language and Terminology</td>
<td>49</td>
<td>33,6</td>
</tr>
<tr>
<td>C8: Applied Classing and Indexing</td>
<td>31</td>
<td>21,2</td>
</tr>
<tr>
<td>C9: Knowledge Organization Environment</td>
<td>48</td>
<td>32,9</td>
</tr>
</tbody>
</table>

*Note: Percentage calculated in relation to the total number of works (146)
Next, the construction of the relationship network between the 43 authors with at least three papers in this period and the main CSKOL categories is presented, based on which, the association between authors and different thematic categories is identified (Figure 1). The intense linking of some authors to specific thematic categories is observed. In the case of C1, the presence of at least four papers by authors Dousa, Gnoli, Green, Olson, Simon, Smiraglia, Tennis and Zeng is seen. In C7, authors such as Priss, Herrera-Viedma, Khoo and Moya-Anégon appear. In C2, with three or more papers, are Broughton, Gnoli, La Barre, Green, Jacobs, Kwásnik, Naumis Peña, Shir and Tennis.

**Fig. 1: Association network between the 43 researchers and CSKOL categories**

The presence of most of the authors in four distinct categories is seen. The authors who appear in seven out of nine categories are: Smiraglia, La Barre and Zeng. The authors present in six categories are: Hadju Barát, Kwásnik, López Huertas, Rafferty and Tennis.

Based on Table 3, depicting the co-occurrence of categories in the 146 papers, it is observed that categories C1, C2, C7 and C9 are those whose thematic relationships are more evident, as they have a higher frequency of co-occurrence, indicating a greater interlocution between them. The co-occurrence between them varied from 7% between C2 and C7 to 14% between C1 and C9. All the categories, except C3 and C4, co-occur with all the others, even if in a less significant way in some cases such as C5. It is seen that C1 is the most related with the others.
Such results are evidence that the authors under analysis have emphasized the theoretical fundamentals in their studies. This happens with more intensity in the relationship of this category with C9, C2 and C7, respectively. Thus, a theoretical deepening is observed in the domain in regard to ethical, cultural, personal and institutional issues, structure development, elements and construction of KO schemes, as well as issues related to semantics, terminology, multilingualism and information recovery systems.

Table 3: Frequency of categories co-occurrence in 146 works

<table>
<thead>
<tr>
<th>Categories</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
<th>C7</th>
<th>C8</th>
<th>C9</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>17</td>
<td>44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td>6</td>
<td>3</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td>5</td>
<td>6</td>
<td>0</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C5</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C6</td>
<td>3</td>
<td>8</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C7</td>
<td>14</td>
<td>10</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C8</td>
<td>3</td>
<td>9</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>C9</td>
<td>20</td>
<td>14</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>13</td>
<td>7</td>
<td>48*</td>
</tr>
</tbody>
</table>

The high incidence rate among categories C2 and C9 suggests that researchers tend to discuss issues such as structuring, planning and the construction of KO systems, while considering social and cultural aspects resulting from the environment that such systems will be serving. This may indicate that even the technical orientation of the field is putting more emphasis on the social aspect.

Conclusions

By means of this study it was possible to identify the most productive authors in KO participating in the International ISKO Conferences, as well as the themes in which the researchers have been publishing. Effectively, the CSKOL categories of greater prominence in the field, as well as the degrees of association between them were identified. The most productive authors were represented by a set of 43 people who presented at least three papers in the last five editions of ISKO international conferences. As seen in Table 1, it was possible to identify that researchers with quantitatively more papers were affiliated to American universities, and that the majority of selected authors were from North-American universities. Through this, it can be affirmed that North-American institutions have more presence in ISKO international conferences.

The synthesis of data presented in Table 2 allowed the identification of thematic tendencies in KO research. It was observed that the category C1, fundamentals of KO has been an important area. The focus on the application of linguistic and terminological aspects to knowledge representation (C7) has been a recurrent research theme. The ethical, cultural and social issues related to the environment in KO (C9), as well as the themes concerning structuring and construction of KO systems (C2), have also been significant areas.

The analysis of the network represented by Figure 1 allowed the visualization of the relationships between authors and thematic categories. The authors and their degree of association were identified for every category. Such analysis was complemented by the results contained in Table 3 about the co-occurrence of categories. It was observed that categories C1, C9, C7 and C2 are more intensely related, C1 being the most linked to the others. This last result reinforces the fact that researchers have focused on the basic
concepts of the domain in order to, not only lay the foundation for their studies, but to also
verify if these conceptions are still capable of explaining social transformations. For this
reason, a high incidence of C9 in relation to the other categories is also observed, as it
represents the social-cultural environments where KO is developed.
To summarize, the construction of theories and models based on the adoption of strict
methodological procedures is shown, with tendencies towards widening the spectrum of
themes related to the social-economical context in which the studies are developed.

References
interlocuções com o universo científico da International Society for Knowledge
Organization (ISKO). Revista Ibero-americana de Ciência da Informação, 1(1): 77-99,
jan./jun.
Hjørland, B. 2002. Domain analysis in information science: eleven approaches – traditional
Hjørland, B. 2003. Fundamentals of knowledge organization. Knowledge Organization,
30(2): 87-11.
Hjørland, B., and Albrechtsen, H. 1995. Toward a new horizon in information science:
Oliveira, E. F. de, Grácio, M. C. C., and Silva, A. C. C. 201). Investigadores de mayor
visibilidad en organización y representación del conocimiento: un estudio desde el
Otte, E., and Rousseau, R. 2002. Social network analysis: a powerful strategy, also for the
Cambridge: Cambridge University Press.
The Evolution of Classification Systems: 
Ontogeny of the UDC

Abstract
To classify is to put things in meaningful groups, but the criteria for doing so can be problematic. Study of evolution of classification includes ontogenetic analysis of change in classification over time. We present an empirical analysis of the [UDC](https://www.ala.org/plc) over the entire period of its development. We demonstrate stability in main classes, with major change driven by 20th century scientific developments. But we also demonstrate a vast increase in the complexity of auxiliaries. This study illustrates an alternative to Tennis' "scheme-versioning" method.

Introduction
Beghtol says that “to classify means to put things into meaningful groups” (2010, 1045). Defining the “things” that are to be classified, determining their attributes and appellations, and drawing inclusion-exclusion boundaries are all dependent on the social construction of knowledge. A classification, being more than a simple tree or simple set of mutually exclusive boxes, is rather more like an evolutionary tree, in that the relative location of species on the tree is affected not only by empirical knowledge of the species but also by the socio-historic character of knowledge in general as it grows unevenly over time. Tennis has written about this evolution in two ways, as subject ontogeny (2002) and as scheme versioning (2007). In his paper on subject ontogeny, Tennis refers to the evolution of classification as the revision of a knowledge organization system ([KOS](https://www.ala.org/plc)) as “a revised scheme absent of any record of what the knowledge landscape looked like before” (2002, 54). His analogy adds the concept of space-time, as the [KOS](https://www.ala.org/plc) which undergoes constant revision becomes perceived as static against a constantly-shifting landscape, which itself appears to have remained static. This process is the phenomenon of instantiation (Smiraglia 2005, 2008), in which information objects are realized at various points along a temporal trajectory. Tennis calls this junction between instantiations of [KOS](https://www.ala.org/plc) and spacetime “the dimensionality of classification” (Tennis 2002, 56). In each case, the failure to recognize earlier or later instantiations, as well as earlier or later knowledge landscapes against which the instantiation is realized, leads to chaotic interpretation. It is this chaotic interpretive line that Tennis refers to as “scheme versioning,” as he offers an empirical approach to locating individual species against these shifting space-time backdrops. The purpose of this paper is to advance understanding of scheme versioning by demonstrating the evolution of a constantly instantiating [KOS](https://www.ala.org/plc), namely the Universal Decimal Classification ([UDC](https://www.ala.org/plc)) system. Elsewhere we have used our evolutionary analysis of the [UDC](https://www.ala.org/plc) as a stable reference system over against the volatility of the knowledge landscape itself represented by the constantly shifting knowledge network in Wikipedia (Akdag Salah et al. 2012; Scharnhorst et al. 2011). In this paper we use the [UDC](https://www.ala.org/plc) as a case study to demonstrate how we have analyzed the instantiating evolutionary tree of the [UDC](https://www.ala.org/plc) over time.
Data and Method

The data that we analyze in this paper stems from different years and formats. We have obtained the first edition of UDC as a printout in French (International Federation for Documentation, 1905). This edition contained only about 400 records, a rough draft of the main classes, and the second classes only. We manually entered the information from the printed edition into an EXCEL file. The rest of our data comes from the UDC archives (the courtesy of UDC Consortium, The Hague). Currently, we have the master reference files of UDC from the years 1994 (61563 records), 1997 (61000 records), 1998 (60685 records), 2005 (66495 records), 2008 (68193 records) and 2009 (68551 records).

In order to analyze the data we decided to apply various visualizations, and chose MagnaView as our visualization software. For this purpose, we have written a Java-based program to convert the UDC Master Reference text file into a tab delimited text file that can be read by MagnaView. In the MRF each record contains one unique UDC number (Strachan and Oomes 1995). “The UDC Master Reference File (MRF) is a database that contains the UDC schedules together with records needed for a The precision of metaphor for dministration, maintenance and archiving”.¹ The process of conversion needed some decisions of how to handle the UDC data. First we parsed the UDC number and cut the string into digits. Hereby, we followed a relative simple method of splitting and counting UDC numbers. The approach is based on the fact that UDC numbers are constructed in a way that directly reflects the class structure, including the fact that the main class number comes first. To investigate the auxiliary usage in UDC we added two extra fields in the MagnaView datafile structure and assigned each record its auxiliary type and extracted the auxiliary part.

Results

Changes in the Main Classes

The UDC has ten main classes (Harper 1954) at the top level of its tree structure. In Figure 1(a) we illustrate the distribution of these ten classes at two different points in time, specifically and respectively 1905 and 1994. Over the century of development two main changes took place: 1) the second level category “[01] Bibliographie” became a part of the main class scheme, and was expanded to include not only library studies, but „Science and Information Organization“ in general. A new addition to this class is „Computer Science“; and 2) the class “[4] Philology. Linguistics” was vacated. This was accomplished by moving „Linguistics“ to “[8] Literature”, and removing the term „Philology“ from the main classes. To capture more fundamental changes between the main classes during the century, we have decided to include „[01] Bibliographie“ as a main class, and highlighted it with the same color of [0] Science and Information Organization. There are two simple reasons for our decision: first, we want to show the expansion of [01] to [0] in numbers, and second, [01] on itself was not a subclass of any other main class in 1905. Otherwise the main class structure of the UDC remained stable, and saw only changes at lower levels. These changes, though not interfering with the main class structure, can be quite severe themselves. The annual issues of the Extensions and Corrections to the UDC list changes such as: entry and exit of UDC numbers, changes in the description of UDC numbers, shifting entries among classes and so on.

Since the launch of the 1993 Master Reference File the revision process is under supervision of the editorial board and its editor in chief. Revisions on specified topics are commissioned and updated tables are submitted to specialists for quality control prior publication. With our visual analysis we highlight some of the traces of those changes echoing the complex editorial process around the UDC in a period of relative stability.

¹ See http://www.udcc.org/mrf.htm for more details.
In relation to 1905, all main classes show an enormous growth, however, if we compare the changes not in pure numbers, but in percentages, we see that [0] has shrunk considerably more than other classes. Actually, except [5] and [6], all the classes’ percentages decreased.

The first editions of UDC attempted to encompass universal knowledge, and that was reflected in the (comparatively) even distribution of the main classes. A comparison of the composition of the UDC between 1994 and 2009 reveals a relative stability in the share of UDC numbers across disciplines (see Figure 1(b)). The modern UDC, as it is obvious in the growth of the proportions, is mainly occupied with natural [5] and applied sciences [6]. This might reflect the increasing importance of science and technology in society, but it might also be a consequence of the development of libraries, and the bias in library collections, for which UDC is mainly used today. What also caught the eyes is the apparent (relative) increase of use of auxiliaries between 1998 and 2005. Obviously, auxiliaries play an important role in the adaptation process of the UDC.

Fig. 2: (a) Distribution of special auxiliaries among the main classes over the years. (b): Percentage of special auxiliaries to main classes, and their changes over time.
Changes in Special Auxiliaries

UDC contains two types of auxiliaries, common and special auxiliaries. Special auxiliaries provide a means to express complexity in proscribed specific classes (thus each is “special” to the class or classes for which it was designated). As we have mentioned before, the usage of special auxiliaries differ among the main classes. UDC contains different types of special auxiliaries used with different connotations.

For the sake of simplicity, we have disregarded these differences, and treated all special auxiliaries as the same. We are more interested in which main classes special auxiliaries are applied and to what extent. Figure 2(a) shows the distribution of them among the main classes, and their changes over the years. As expected, main class [6] has the most special auxiliaries, even though their number is on the decrease from 9613 in 1994 to 9442 in 2009. Here the rate of decrease follows a soft curve, in contrast to a few jumps we observe in other main classes. For example, class [2], i.e. Religion, has an unprecedented change in the special auxiliary numbers from 1998 to 2008: from 239 records to 2168 records. If we consider the total record amount class [2] covers, we see that this class has seen a substantial change over time: In 1998, [2] contained only 935 records, which has risen to 2419 records in 2008. The transformation of special auxiliaries here is paramount; in 1998 only 14.87% of the class [2] was occupied by special auxiliaries, whereas in 2008 this number has risen up to 89.6%.

Similarly, we observe sharp fluctuations among the classes [5], [8] and [9]. To understand the importance of these, it is best to compare the changes in connection with the changes in the main classes themselves. Figure 2(b) shows the changes of special auxiliaries, but this time, the y-axis of the chart reflects the percentage of the special auxiliary record numbers to the total number of any given main class. Thus we see that after 2008, the main class [2] is been transformed almost entirely; now % 90.07 of it is only special auxiliaries. The main classes [8] and [7] also have a considerable amount of special auxiliaries (50.7% and 40.48% respectively), whereas in the main class [9] the use of special auxiliaries has been reduced from 44.58% in 1998 to 6.75 in 2009.

The impression we get from the Figure 2(a) is that the most special auxiliaries appear in the main class [6]. When we speak of the number of records a given class covers, this impression is correct. However, if we want to see the percentage of each class occupied by special auxiliaries, we must consult the Figure 2(b).

Changes in Common Auxiliaries

Common auxiliaries are used to express language, form or cultural origin; they are called “common” because they can be used with any number where appropriate (thus they are common to the entire UDC). Common auxiliaries in the UDC are on the increase. As also indicated in Figure 1(b), in 1998 the number of common auxiliary records was 6812, whereas today, this number has reached a total of 13562. So, the use of auxiliaries grows over-proportional in the UDC evolution, especially the common auxiliaries. In comparison to this sharp increase, all other fluctuations among the main classes seem to be trivial.

In order to understand why common auxiliaries occupy so much more space in UDC, and how they change the structure of the UDC tree itself, we need to further analyze this phenomenon. Figure 3(a) displays the distribution of common auxiliaries over time. Since our data stems from the Main Reference File, and does not reflect the actual usage of UDC, common auxiliaries such as „a“ and „b“ have only one record, and that is the description of how these auxiliaries should be used. These auxiliaries are above all for combining different UDC main records, i.e. records from the Main Table, and thus in our analysis we cannot trace their effects. A similar common auxiliary is the type „h“, which incorporates
non-UDC sources into UDC records. This auxiliary also come to life only in actual usage, and does have only one record in all the MRFs.

The rest of the common auxiliaries need to be assessed further: Among these, "e", the auxiliary of the place, is the most used one, as well as the one with the highest increase in numbers. The second most used common auxiliary is the type "c", auxiliary of the language. However, the second auxiliary that has a sharp increase is the auxiliary of persons and materials, i.e. the auxiliary "k". The auxiliary of form (type "d") remains stable, and the auxiliary of time (type "g") shows only a small increase. Thus, we can conclude that the main changes in the record numbers of the MRF files from 1993 to 2009 have taken place among auxiliaries, and among them, the reason of this sharp increase comes from the additions to the auxiliary "e" and "k".

**Figure 3:** (a) Distribution of common auxiliaries over the years. (b) Changes in the record number of main classes from 1993 to 2009

### Conclusions

With this analysis we examine some features of the dynamics of a knowledge organization system in time. Our explorations are of importance along different dimensions: theoretical debates inside of knowledge organization and information sciences; the UDC consortium; and aspirations to better understand and manage complex knowledge spaces across many scientific disciplines.

### Ontogenetic analysis

Tennis (2007) proposed a specific schema for ontogenetic analysis of individual concepts as they evolve over time in a given KOS. Our approach has been different, in that we have chosen to analyze the evolving structure of the UDC against the socio-cultural knowledge landscape of the 20th century in which it developed.

For the most part, change in the UDC reflects well the growth of knowledge over the past century. Interestingly, we have observed a vast increase in the auxiliaries. This includes the enumeration of peoples and places in the common auxiliaries, reflecting increased cultural sensitivity in the UDC, but also special auxiliaries which increased dramatically in number in applied sciences and religion. This is a shift in what Tennis called “dimensionality,” as the UDC grew not in its extension, but rather in its intension, as the branches of the initial tree were expanded with auxiliaries. Here, Beghtol’s inclusion-exclusion boundaries shift
in dimension across spacetime. Hence we understand there is not at any point a single UDC, but rather, at every point, there is the UDC in spacetime.

Our results demonstrate the efficacy of this analytical approach to ontogenetic analysis. We believe that this approach, combined with Tennis’ suggested approach to scheme versioning provides a useful methodological triangulation that incorporates empirical knowledge about the evolving instantiating, living, UDC.

The UDC Consortium

Not surprisingly with a history of more than 110 years, the UDC has gone through many transformations in itself, but also in the ways it is managed. By 1980s, UDC had many different versions in various languages, and its full version exceeded 200,000 records. To produce editions of the UDC in different languages which can be applied in collections around the globe is one focus point of the editorial board. Another one is of course the adaptation of the UDC to the growing and changing body of knowledge. A great attention and care is devoted to first make this changes and with them keep the UDC as a living knowledge order system (and not a historic one); but second also to do it with care to guarantee a certain continuity so that collections don’t have to re-index their holdings all over the time. Not surprisingly, the growth and change in the UDC did not always follow a strict strategy with some problematic results, as the editorial board pointed out in several publications themselves. For example, the original 19th century knowledge-base had overlapping concepts from the 20th century. As one solution a medium version is generated by a selected editorial board (Slavic et. al., 2008). This was formed in 1989 from the 1985 English edition, which was already in machine-readable format (McIlwaine 1997). The resulting updated Master Reference File was launched in 1993.

Our analysis traces some of the changes in a period the editorial regime of the UDC was relative stable. However, the changes we reported can only be properly understood knowing the context and history of editorial changes in the UDC. Sudden changes in records in some classes can be related to specific concerted action of their profoundly renewing. Auxiliaries are used to avoid repetition of subclasses through different main classes, but seem also have to be used to add diversity and flexibility without disturbing the overall structure to much. The different editions of the “Guidelines to the UDC” and further publications from the consortium (e.g., Riesthuis 1998) and others contain such context information. Yet a comprehensive history of the UDC including elements of statistical and visual analytics is pending. Nevertheless, we believe that this kind of statistical analysis can also trigger further debates about the UDC as complex KOS. (Scharnhorst et al 2012)

Evolving Knowledge Spaces

The UDC as a complex KOS is rather a language to describe objects in a controlled way than a hierarchical tree which allows to (co-)allocate objects. The possibilities to combine UDC numbers directly but also their indirect coupling via shared auxiliaries makes the UDC a complex network. (Barabasi 2002, Scharnhorst 2003) With this paper we only looked at the surface of what is possible in terms of analyzing the evolving network of the UDC. Moreover, we analyzed the UDC only in terms of its Master Reference File appearance. One could say that this is the “designed” side of the UDC – or the UDC master plan. The application of the UDC in libraries and museums creates another rich set of UDC expression in collections. In other words, via a shared KOS as the UDC we can analyze the evolution of knowledge as recorded in space-time specific objects in collections, which are also divers in time and space. Insights in the dynamics of knowledge production and its organization can help us to better understand and navigate through large-scale knowledge spaces as available via the web.

2 Personal communication Gerhardt Riesthuis.
References


Facets and Fugit Tempus: Considering Time’s Effect on Faceted Classification Schemes

Abstract
Describes the effect of scheme change on the semantics in faceted classification. Two types of change are identified: ecological change and lexical change. Examples from different editions of the Colon Classification are used to illustrate change.

Introduction
S. R. Ranganathan sought a solution to the enumerative classification problem. That is, if we have an ever-expanding universe of knowledge, how do we arrange new subjects in relation to those already in our scheme for classification? His proposal was to break subjects apart. These parts he called facets. The classifier then recombines them using principles and postulates into a meaningful and systematic order. This allowed the classifier to add facets (parts of a subject) to the schedule and then follow the postulates and principles so that this subject (with a part newly added and represented in a class number) would fall into place in the hierarchy and the systematic arrangement of classes that is more meaningful and therefore more helpful to users. Classification that uses both facets and postulates, and principles is analytico-synthetic classification (Ranganathan, 1967).

Though analytico-synthetic classification is successful at addressing the challenge of adding new knowledge to a schedule, there remain at least two challenges: (1) the potential change in meaning of extant facets with the addition of new facets, and (2) the change in definitions of a single term over time. These are interrelated problems, but separable. When facets change their meaning because of the addition of new facets we have an ecological change. When we change the definition of a term because society has recast the term with a different intension, then we observe a lexical change.

Lexical Change in Analytico-Synthetic Schemes for Classification
Whereas the addition of new subjects is mechanically and theoretically fixed in Ranganathan’s work, he did not account for changes in meaning (both extension and intension) of facets once they are established in a particular scheme. For example, CIVIL ENGINEERING as a subject has changed over time. If we were editing an enumerative classification scheme we might subdivide CIVIL ENGINEERING into parts to represent a more refined understanding of the topic in subsequent editions for the scheme. However, we are not asked to do this in an analytico-synthetic scheme for classification. We assume a constant meaning. An editor might also erase an older meaning in a new edition of an enumerative classification scheme, retaining the word, but repositioning it in a different main class. This is what has happens with EUGENICS in the DDC (Tennis, Thornton, Filer, 2012). It is no longer possible to define EUGENICS as a biological science in DDC. We are forced, by the scheme to define anything written on eugenics to be applied science, social science, philosophy, or history. This kind of change is lexical change.

Ecological Change
Above we mentioned the addition of facets to an analytico-synthetic scheme would not assume a change of meaning of extant facets. In fact, we have to make such an assumption or else we undermine the value of facets in this kind of scheme. We have to assume that the meaning of a classmark is constant in relation to potentially new classmarks generated
from the addition of facets to the scheme for classification. However, if we do recognize that the meaning of CIVIL ENGINEERING has changed over time then we might add new facets to the scheme to represent the emerging concepts. This in turn, empties out the meaning of CIVIL ENGINEERING. It will re-contextualize, for both the classifier and the user. Up to the point of adding new facets, all things about CIVIL ENGINEERING goes in one facet, and after the new facets fewer things go there.

**The Problem**

The question for the editor of an analytico-synthetic classification scheme is then, the same as that of the editor of an enumerative one. Both ask, what should be done to revise extant topics in a scheme for classification if their meaning shifts? The problem for the editor of an analytico-synthetic classification, however, is complicated by the fact that new meanings introduced as warrant are not meant to revise a facet, but rather to add a facet. This is a problem of the metaphor of space in classification schemes. Enumerative schemes have a discrete number of places for concepts at any one time in any single edition. An analytico-synthetic scheme for classification has a combinatoric number only once. As soon as there is a change in meaning, we must recombine the space and meaning metaphor and rerun the combinatoric calculation. This breaks the semantic line from one edition of the analytico-synthetic scheme for classification to another. Thus, the shelf order, derived from the meaning of the facets combined into a classmark, loses meaning. Since the problem here is the break in meaning from one edition to another, we must study the anatomy of the problem and provide an amelioration. What is the nature of this break and how do we correct for it in our systems? This is a challenging set of questions.

**The Challenge**

When enumerative classification schemes change we make a decision to reinvent the classification. In so doing, we issue a new edition of all the possible classes. If classes change, the line between the classes of the old edition and the new edition are broken in particular and documented ways (Tennis, 2005 and 2007). The old class continues, is erased, or is replaced by a new class or constellation of classes.

However analytico-synthetic schemes for classification are different. The line between editions is broken in a different way. The old facet is not always erased, even though other facets appear that complement its meaning. We must document how this change occurs in this particular kind of system. First we have to account for the lexical and ecological change as described above. Then we have to account for the interaction between scheme change and postulates and principles. This is a unique challenge of studying scheme change in an analytico-synthetic scheme for classification.

**Method of Analysis**

Using extant work on scheme change (Tennis, 2007; Tennis and Sutton 2008; Tennis 2011a, Tennis, Thornton, and Filer, 2012) as a baseline, this paper examines the processes involved in the change of an analytico-synthetic scheme for classification. Key to this investigation is a clear understanding of the relationship between postulates and principles in classification and the structure of the schedules (as the manifestation of the scheme). In order to comprehend this interaction, we use extant examples of the Colon Classification as a case study.

**Findings**

When we examine facets and isolates in the various editions of the Colon Classification we can see the effects of both ecological change and lexical change. Starting from the third edition (Ranganathan, 1950) when postulates and principles were mature in Ranganathan’s
thought, we followed GENETICS, ALTERNATION OF GENERATION, and ONTOGENY. We can see the lexical change reflected in the headings that go with the isolate numbers. With GENETICS we see an erasure of a refined meaning because the editors drop PHYLOGENY as a qualifying term. See Table 1.

Table 1: Biology Isolate Numbers in a Sample of Editions of the Colon Classification

<table>
<thead>
<tr>
<th>Date and Edition Number</th>
<th>Isolate Number</th>
<th>Isolate Heading</th>
</tr>
</thead>
<tbody>
<tr>
<td>1950 3rd Edition</td>
<td>6</td>
<td>Genetics, phylogeny</td>
</tr>
</tbody>
</table>

We can infer that phylogeny is not, according to the editors, the only way to conceptualization and describe genetics by 1987. And this tracks the change in the science. So we see how the line breaks here. This is lexical change.

Table 2 shows the ecological change in the same area of the Colon schedules.

Table 2: Sample Isolate Notation in Four Editions of Colon Classification

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Ontogeny</td>
<td>Ontogeny</td>
<td>Development (ontogeny)</td>
<td>Development (Ontogenesis)</td>
</tr>
<tr>
<td>71</td>
<td>Fertilisation</td>
<td>Fertilisation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>715</td>
<td>Artificial fertilisation</td>
<td>Artificial fertilisation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7192</td>
<td>Twins</td>
<td>Twin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>Germination</td>
<td>Germination</td>
<td></td>
<td></td>
</tr>
<tr>
<td>73</td>
<td>Embryology</td>
<td>Embryology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>Growth after birth</td>
<td>Growth after birth</td>
<td></td>
<td></td>
</tr>
<tr>
<td>751</td>
<td>New born</td>
<td>New born</td>
<td></td>
<td></td>
</tr>
<tr>
<td>752</td>
<td>Toddler</td>
<td>Toddler</td>
<td></td>
<td></td>
</tr>
<tr>
<td>755</td>
<td>Infant</td>
<td>Infant</td>
<td></td>
<td></td>
</tr>
<tr>
<td>76</td>
<td>Pre-adolescent</td>
<td>Pre-adolescent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>77</td>
<td>Adolescent</td>
<td>Adolescent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>78</td>
<td>Old age</td>
<td>Old age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>791</td>
<td>Death</td>
<td>Death</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The 7th Edition has this note: “Divisions as for "7 Development" in the Schedule of (1MP1) isolates for "L Medicine" in Chap EL. Schedule of (1Mmt1) Substance isolates. Divisions as in the schedule of (1P1) isolates "E1 General Chemistry" Chap EE. Schedule (1E) Action isolates (For use also in (1E) of GT, GV, I, and K” (Gopinath, 1987). This directs the classifier to another part of the schedule.

Table 2 is a clear example of ecological change. We removed isolate numbers from this subject, removing the expressivity. From this we can see how the line is broken from one edition to another.

The final example from this part of the Colon Classification, ALTERNATION OF GENERATION, has a single change in the 7th Edition. We add a synonym. ALTERNATION OF GENERATION (METAGENSIS). It is not clear what effect this has on the semantics. It could be either a lexical or ecological change if METAGENSIS was defined or used elsewhere.

Constraints to Our Work

We are constrained by our definition of meaning. We make assumptions about the philosophy of language and meaning when we invest in the design and implementation of
an analytico-synthetic classification scheme. Knowledge organization literature has taken to
task positivistic and referential theories of meaning. And leans to more pragmatic and
critical prescriptions of how indexing languages (of which classification is a type) work. In
this context we can observe an axis with poles. At one end we have Svenonius”s
construction of referential semantics. Her example is Wittgenstein”s picture theory of
meaning described in the *Tractatus*. She says, “The basic tenet of the picture theory is that
the extensional meaning of a word is its referent. For example, the extensional meaning of
the word „butterfly“ is the set of all past, present, and future butterflies. Words whose
referents are things in the real world can be taught by ostensive definition, simply by
pointing to their referents,” (Svenonius, 2004). The value of this theory of meaning is that
once identified and named, the concepts do not change. Further, the relationships between
these concepts are of two kinds: contextual and permanent (independent of context). Some
knowledge organization researchers set as a design desiderata that only permanent
relationships be admitted in the indexing language, and they assumed the permanence of
concepts because of referential semantics (cf., Bernier, 1968). Svenonius finds the picture
theory lacking, as do other thinkers in knowledge organization, and the trend in the field is
to assume some sort of pragmatic view of meaning (cf., Svenonius 2004, Mai, 2004). It
seem that we have to consider context in scheme change, because we are dealing with
multiple schemes (editions over time), a change in retrieval sets (documents classed in
1920 vs. 1980), and our assumptions about literary warrant both contemporary and
historical. Yet we are concerned with a theory of meaning that will allow for the
preservation of meaning of retrieval sets over time.

A retrieval set is the set of documents assigned to a class. As new materials are attached
to a class the extension of that class has the potential to shift, thereby shifting the
interpreted meaning of the class. The design of analytico-synthetic schemes for
classification does not currently account for this, and we must weigh the structural scheme
changes (addition and/or deletion of facets) against this more literature-based definition of
meaning. This also constrains our meaning.

**Toward Conclusions**

We are able to show at this point that there are three levels of analysis of change in an
analytico-synthetic scheme for classification. We must (1) document change of facets in
relation to each other and (2) change in potential subjects constructed from combining
facets. Finally, (3) we must document change in the explicit relationship structure over time
in the Colon Classification, our example.

In the case of our biological examples above we can see deletion of facets and pointing to
other areas in the schedule for classification. This is the first level of analysis. The change
in the potential subjects that can be represented seems straightforward. By 1960 we cannot
use G7192 for twin ontogeny. We have to start with G7 (ontogeny/development), and then
add a connecting digit without direction. In 1987 we have guidance on this. So there are at
least three potential notations for the ontogeny of twins in the life of the Colon
Classification. This is the second level of analysis. The third level requires us to link these
three notations. This third level is ongoing research and will be reported elsewhere.

Change in an analytico-synthetic classification scheme is much more a web of
connections, and the complexity of mapping these changes is quite large. However, there is
value in the exploration of this complexity for our understanding of both improving
systems, and revisiting our theory of meaning in knowledge organization. We are forced to
consider which contexts matter and which do not in our pragmatic thinking when we
examine facets and how *fugit tempus* – time flies.
References
Sharada, B. A. – Central Institute of Indian Languages, Mysore

Ranganathan’s Colon Classification:
Kannada-English Version ‘dwibindu vargiikaraNa’

Abstract
“dwibindu vargiikaraNa” (ದ್ವಿಬಿಂದು ವರ್ಗೀಕರಣ) is the Kannada rendering of the revised Colon Classification, 7th Edition, intended essentially for the classification of macro documents. This paper discusses the planning, preparation, and features of Colon Classification (CC) in Kannada, one of the major Indian languages as well as the Official Language of Karnataka, and uploading the CC on the web. Linguistic issues related to the Kannada rendering are discussed with possible solutions. It creates facilities in the field of Indexing Language (IL) to prepare products such as, Subject Heading List, Information Retrieval Thesaurus, and creation of subject glossaries or updating the available subject dictionaries in Kannada.

Need for ILs in Indian Languages
India is linguistically rich with 1,652 mother tongues of which 22 are Scheduled Languages (and 100 are non-scheduled languages) included in the Constitution of India. However, there is a paucity of indexing languages (IL) in Indian languages (Sharada, 2010). Most of the existing IL are in English and a few other languages like Chinese, French, German and Italian. Realizing the importance of IL in Indian languages, M.A. Gopinath had written in the Introduction of the 7th edition of the Colon Classification (CC) that he hoped to develop the scheme in Indian languages as well, in collaboration with other organizations (Gopinath, 1989). The history of documents published in regional languages of India goes back to the early 19th century. For instance, the first book in Kannada was published in 1817 (Wikipedia). There are innumerable concepts native to a particular region, language and culture which are not found in the English language schedules. Tools to properly organize documents on such concepts in a user-friendly sequence are lacking and therefore, there is a need to prepare IL in Indian Languages.

CC in Kannada and other Indian Languages
The first rendering of the CC 7th English Edition into an Indian language is “dwibindu vargiikaraNa” prepared and published through the collaborative efforts of Sarada Ranganathan Endowment for Library Science (SRELS), Bangalore and the Central Institute of Indian Languages, Mysore (CIIL) in 2010. Udaya Narayana Singh, former Director of CIIL, in his foreword to this volume, states that:

“On the event of Kannada getting the Classical status, with its richness of literature in all walks of life, what was lacking was the tool to properly organize the universe of knowledge. Hence, as a test case, Kannada was accorded priority.”

This has opened up opportunities to develop several information retrieval tools in Kannada. This is the first bilingual classification schedule adequately updated and flexible enough to accommodate any number of new concepts.

Linguistic Features Across Language Families
A major linguistic discovery of the 20th Century relating to Indian languages is the identification of common linguistic features across different language families. Indian languages belong to different language families but do share common linguistic features, hence India is referred to as a „linguistic area” (Emeneau, 1956). This sharing of linguistic features across the language families was facilitated by their coexistence for centuries together, and by the continuing interaction between the people speaking these languages.
The language families are: Indo-Aryan (Hindi, Bangla, Marathi, etc.), Dravidian (Kannada, Malayalam, Tamil, Telugu, etc.), Tibeto-Burman (Khasi, etc.), Astro-Asiatic (Santhali, Mundari, etc.) and Sino-Tibetan (Bhutanese, etc.). A 19th Century missionary to India, Rev. William Campbell wrote (1839):

"Whatever may be the difference in the languages, they all belong to the same great family; similar laws regulate the idiom, construction, style, and various kinds of composition, which prevail in the dialects of the north and the south; when you describe one art of India, you have, in many respects, described the whole; the manners, the customs, and the habits of the people, with trifling variations, correspond from Cape Comorin to the Himalayas; and their superstition, in all its great lineaments, is exactly the same. Whether, therefore, their present literature was originally written in Sanskrit, or in some other languages, the Vedas, the Shastras, the Purans, and all their classical writings are to be found in all the principal tongues of India, and are as well understood in the one as in the other"

Languages within a group also share a number of common elements. Due to shared linguistic features like vocabulary, and grammatical elements, and common patterns followed in the creation of technical terminologies by the Centre for Scientific and Technical Terminology across the languages in the post-independence India, rendering the same into other Indian languages (with one or two exceptions) should not be a difficult task.

Some of the shared linguistic features across language families are as follows:
1. Presence of a series of retroflex consonants that contrast with dentals sounds [t/ D].
2. Two to three degrees of 'you'. [tu „you” (sg), tum „you” (pl) and aap (hon) in Hindi and niinu „you” (sg), niivu „you”(pl) and taavu (hon” in Kannada)
3. Widespread lexical borrowing.(samaya, kaala, aadeesha)
4. Presence of echo word constructions and onomatopoeic forms. (gaD baD, daDa)
5. Reduplication process of different grammatical categories such as nouns, verbs, adjectives, adverbs, etc. (baDe baDe, doDDa doDDa)
6. Compound verb forms.
7. Conjunctive particle. (uu „also”)
8. Sentence structure - flexibility of word sequence though finite verb usually comes in the last position. (avaru manege baruttaare, vah ghar aayeegi)

**Relevance of CC**
S. R. Ranganathan’s Colon Classification (Ranganathan, 1963) based on his theory of freely faceted classification has a modular structure and is sufficiently flexible to enable the addition of new concepts without the structure being seriously disturbed (Neelameghan, A., 2000).

**Planning and Preparation**
Data was collected from cc 7th Edition and broad translation was done. Each descriptor in the schedule in Kannada was rendered after checking subject dictionaries/glossaries available in many forms such as print, manuscript, and digital. Subject experts with proficiency both in a particular subject as well as the Kannada language, were involved to assist in this venture. Chapter headings, representation of Basic subjects within a main subject, notation, descriptors, expression of rules (example: Divide by CD), etc, were planned before preparing the schedules so as to ensure consistency and uniformity in the entire schedule.

Considering that many developments had occurred since the 7th edition of cc was published, the schedule of Basic subjects in the English version was revised. New concepts were introduced without disturbing the existing ones wherever necessary. For example: BX Astronomy: (Ranganathan, 2010, p.167):
The book has six chapters. Chapters 1 to 3 provide an introduction to CC, its postulates and principles. Chapter 4 lists traditional main subjects, schedule of basic subjects, and schedules for languages and common isolates such as Time, Space, Energy, Matter Property and Personality facets. Chapter 5 provides schedules of special isolates for various Basic classes. Chapter 6 is the Index and the list of books referred.

**Common Isolates**

In the schedule of languages, due importance is given to Indian languages, with special reference to the languages mentioned in the 8th Schedule of the Constitution of India. Similarly in Space schedule, the districts in the state of Karnataka are fully listed. Common schedules for Matter Property and Energy Isolates have been expanded.

**Special Isolates**

The „E Chemistry“ schedule of Inorganic elements, 16 groups are listed as against 8 in the English edition. In „I Botany“ and „K Zoology“ schedules, based on International Code of Botanical Nomenclature (ICBN) and International Code of Zoological Nomenclature (ICZN) respectively scientific names in Botany and Zoology are Latinized. Its intent is that, each taxonomic group of plants has only one name accepted world wide. Many plants of Indian origin are included. Arts and crafts related to Karnataka are also included. Newly added subjects are:

- 5.9B ಗಣಕವಿಜ್ಞಾನ - Computer Science
- 5.AB ಪರಿಸರವಿಜ್ಞಾನ - Environmental Sciences
- 5.AC ಮಗಲೆೈವಿಜ್ಞಾನ - Surface Science
- 5.3X ಪುಸತಕಪರಕಟಣೆ - Book Publishing
- 5.LZ ಕ್ರಗಡೆಗಳು – Sports

**Index**

Many concepts at micro level not included in the schedules may be found in the index. The present work is for classifying macro-documents. Instead of elaborating some of the concepts in the main schedule, they have been included in the alphabetical index.

**Special Features**

CC is facet based and recognizes five Fundamental Categories (FC) as mentioned in Table 1.

<table>
<thead>
<tr>
<th>Name of FC</th>
<th>Indicator Digit</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personality</td>
<td>,</td>
<td>P</td>
</tr>
<tr>
<td>Matter Material</td>
<td>;</td>
<td>MM</td>
</tr>
<tr>
<td>Matter Property</td>
<td>;</td>
<td>MP</td>
</tr>
<tr>
<td>Energy</td>
<td>:</td>
<td>E</td>
</tr>
<tr>
<td>Space</td>
<td>.</td>
<td>S</td>
</tr>
<tr>
<td>Time</td>
<td>“</td>
<td>T</td>
</tr>
</tbody>
</table>

The Facet formula BS, [P]; [MP];[E];[S];[T] is applicable to all the schedules. The abbreviations for the five FCs mentioned in the above Table 1 and the five devices
mentioned below are retained as they are, since these devices are more familiarly known in the abbreviated form in English.

- **CD** Chronological Device
- **GD** Geographical Device
- **SD** Subject Device
- **AD** Alphabetical Device
- **ED** Enumeration Device

Throughout the schedule these devices are used as mentioned below:

(AD ಅಂಕೆಯ ಭಾಗವನ್ನು - Division by AD)

{ದೇಶಕ್ಕೆ ಭಾಗವನ್ನು - Example}

zG ನಿಂದಪಡೆಯ ಭಾಗವನ್ನು - Gandhiana

### Linguistic Issues Related to Rendering in Kannada

**a)** For a single English descriptor different terms are found in different dictionaries. While taking into consideration the context, one has to depend on the entry in CC schedule based on the subject.

For example: Male – *gandu, ganDasu, purusha, manushya*

Though all the four terms are in use, the word most suitable to the subject has to be chosen for representation.

**b)** Since a few descriptors do not have exact Kannada equivalents, they have been adopted as they are in English (transliteration). Schedules for Space Isolates, Botany and Zoology are good examples. The major problem with transliteration is for the words beginning with Vowels such as E, I and O and consonants – J, C and K. In textual transliterations it may depend upon the context. As we are considering only the concepts, based on the phonetics, equivalent terminology is provided. For the pronunciation, standard dictionary and the internet site http://www.answers.com is referred.

For the term Engineering:

1. ಇಂಜಿನಿಯರಿಂಗ್ *inginiyaring* [University of Mysore English – Kannada Dictionary]
2. ಇಂಜಿನಿಯರಿಂಗ್ *enginiyaring* [Vijnaana Kosha]

Both are in use, but we have to adopt one. The latter is chosen since, in the transliteration, the letter “e” represents Kannada letter “ಆ” [ex: *ettara ಎತ್ತರ*] and “i” represents “ಇ” [ex: *ಇಂಡಿಯ - india*].

In some cases, borrowed words from Indo-Aryan languages into Dravidian with localization are chosen. For example: Manager – *nirvaahak, prabandhak* are the Indo-Aryan equivalent. It is transformed into Kannada by adding an inflection „a“ and the Kannada word is *nirvaahaka or prabandhaka*.

**c)** Derivational implications

The term “Chemistry” has two renderings ರಾಸಾಯನಿಕ ಶಾಸ್ತ್ರ *raasaayanika saastra* and ರಾಸಾಯನ *rasaayana Vijnaana* in two different dictionaries. The latter is determined to be appropriate since, in Kannada, if “ಅ” (ika) prataya /inflection is added to a root word, the first syllable of the word becomes long. Several examples are there for this group.

ಹೊಸ ಸಾಮಾಜ *samaaja (Society) - ಸಾಮಾಜಿಕ suamaajika (Sociological)*
Web Interface

In December 2011, a website on dwibindu vargiikaraNa was developed (accessible at http://www.classicalkannada.org.) Here, the full text of preliminary pages and first three chapters can be found. For the schedules of common isolates, and special isolates sample page link is provided including the index.

Uses of DwibinduvargiikaraNa

It is helpful to library professionals who are educated in Kannada medium throughout their educational career.

1. There are several concepts native to a particular region and its culture. Folk arts of Karnataka are a good example.

2. The elaborate alphabetical index is useful in formulating subject headings in Kannada, as the basis for the development of an information retrieval thesaurus as also lexical tools similar to WordNet, technical glossary/dictionary, etc

3. The present work "dwibindu vargiikarana" can be used as a base for translating CC into other Indian languages

4. Subject experts who helped in this work have also found it useful for them to teach their subject in Kannada. The hierarchical arrangement of concepts is very useful to them.

Future Agenda

1) Based on the dwibinduvargiikarana, essential information processing and retrieval tools in Kannada, such as thesauri (IRT) and lists of subject headings are being planned. A bilingual thesaurus (Kannada- English) is in progress. As a test case the discipline “Education” has been selected. The related thesauri available online is being used for English and are rendered to Kannada with class numbers from CC.

The need for a list of subject headings in Kannada was much felt when - library retro conversion work was in progress. The cataloguing of books in Indian languages was done in both the language of the item and in English except for the subject heading field which had to be only in English due to the non-availability of such tools in Indian languages.

2) Natural Language Processing (NLP) application

The development and applications of NLP in Indian languages are at various stages of development. In a recent survey of publications between 2000 and 2010, 297 papers on development and application of NLP in information retrieval in Indian languages were found of which only 14 papers were on Kannada which stands fifth in the rank list (Sharada 2012).

Earlier studies in English in this area tried to find the suitable grammar worked out for natural language. It is useful to explore the feasibility of keeping the five fundamental categories as the delimiters to develop parsers for Indian languages. Separate files need to be prepared for Basic Subject (BS) and the five FCS with an algorithm. After developing this, the schedule can be used online with the flexibility to instantly incorporate new developments in the respective disciplines.
References


Abstract

There are close to a billion websites on the Internet with approximately 400 million users worldwide [www.internetworldstats.com]. People go to websites for a wide variety of different information tasks, from finding a restaurant to serious research. Many of the difficulties with searching the Web, as it is structured currently, can be attributed to increases to scale. The content of the Web is now so large that we only have a rough estimate of the number of sites and the range of information is extremely diverse, from blogs and photos to research articles and news videos.

Introduction

While much of the publicly available printed material has been classified using prescribed methods and trained personnel, this approach falters in the context of the Web. Not only because of the scale of the knowledge base but also from the recent pressure by users to complement or even replace the authority for classification from classification schemes and controlled vocabularies to the largely unstructured approach of social classifications by users. Traditional knowledge organizations have been designed to capture the semantics of the content for the purpose of access to that information, while social classification, such as tagging, has been designed to capture user sentiment about content for informing a community about the content.

The effective creation of a knowledge organization and the retrieval of that knowledge is, of course, deeply rooted in our ability to consistently label and organize content, whether by controlled vocabularies, uncontrolled vocabularies, hierarchical and faceted classification schemes, inverted lists of words in documents, and whether done by professionals, algorithms, or users.

In this paper, we address the emergent role of social classification, specifically user driven tagging, in building knowledge organizations of Web content. The aggregation of user assigned labels or tags of content (from photos to research articles) forms a loose taxonomy, a folksonomy (Vander Wal; 2005), that describes the content in order to provide access to the content for other users. Tag clouds, first made popular by Flickr [www.flickr.com] to label photos uploaded by users, are the most prevalent visualization technique for collaboratively created tags. Figure 1 illustrates a tag cloud where the tags are presented in alphabetical order and the size reflects the relative frequency of the tags in the collection. Another variation of a tag cloud, shown in Figure 2, displays semantically similar tags as a cluster, in this case linearly, within the tag cloud.

The growth in the use of user generated tags to organize and to provide access to information or artifacts on the Web raises several important questions for both retrieval and classification of the information; the quality of the resultant indexing, the consistency of indexing across data, the stability of indexing over time, the effectiveness of retrieval using the resulting indexing, and the relationship of conventional classification to social classification.
Social classifications build collaborative vocabularies dynamically as users describe their content and the content of others. This raises concerns over whether the resultant communities, in fact, share a vocabulary, which is essential to effective retrieval, and, indeed, whether a community vocabulary does emerge, or is stable over time. A set of three studies using CiteULike and Connotea (Kipp & Campbell, 2007) showed that consensus on vocabulary does emerge and further that stability in that vocabulary remains, at least after the two years of the study. Recent research (Robu et al, 2009) showed that in large collections of user tags common structures and categorizations emerge and Wetzker et al. (2010) showed that individual tag vocabularies (personomies) could be translated into shared vocabularies.
The argument is not that social tagging will, or should, replace well considered classification schemes but rather that given a sufficiently large set of such tags in a domain of interest a classification vocabulary and a form of knowledge organization will emerge. Interestingly, user created tags add a different perspective to classifications, and can be seen as complementary to semantic content. For example, social classifications often include concepts of intended action and sentiment along with semantic content. Intended actions, such as “see this” or “read later”, differ from the semantic goals of traditional classification labels (Kipp & Campbell, 2007). In social classification schemes a user may include judgments, personal tags, task related actions, or signals to collaborators in tags as well as content references.

The identification of relevant information on the Web has two phases; the identification of a web site, typically by a keyword search, followed by the identification of relevant information within that web site, typically by following links. Much of the research on Web access has focused on the first phase, the identification of sites, and less on the second phase, the identification of relevant information within the website (Lazar et al., 2003; Tausher & Greenburg, 1997) and navigation within a website remains one of the main causes of user frustration on the Internet (Lazar et al., 2003).

Websites typically incorporate navigation tools to help users find relevant information within the site; menus, lists, tag clouds, and embedded hyperlinks. Although these tools are easily recognized and heavily used, they remain problematic for users when the categories chosen for instantiating the menus or links, created by the author of the site, do not match the context or knowledge organization as understood by the user. Good menus, good lists, and good links depend on the quality of knowledge organization as defined by the categorization and the knowledge structure describing the content of the site.

In a recent study we explored how users use multiple navigation tools (menus, search and tag clouds) to perform information finding tasks on four different websites. Figure 3 illustrates a website featuring the three navigation tools.

Fig. 3: Three Navigation Tools on Many Eyes [www-958.ibm.com]
In addition, the study compared two variations of tag clouds, one where the tags are created by a single user and the other where the tags are created by a community of users. The websites along with the task and the type of tag clouds are illustrated in Table 1.

**Table 1: Websites, Tasks and the Type of Tag Cloud**

<table>
<thead>
<tr>
<th>Website</th>
<th>Description</th>
<th>Type of Tag Cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professional on the Web</td>
<td>Companies/professionals create their profile and also create tags for their own services. Task: <em>Find two companies that you would consider hiring to redesign a website.</em></td>
<td>Community Driven</td>
</tr>
<tr>
<td>Many Eyes</td>
<td>Users create/share visualizations. Users assign tags to their own visualizations. Task: <em>Find two visualizations beneficial for your project on alcohol consumption and its effects.</em></td>
<td>Community Driven</td>
</tr>
<tr>
<td>Web Designer Wall</td>
<td>A blog of web design ideas, trends and tutorials. Maintained by a single author who tags the blog entries. Task: <em>Find two posts on how to design a website for mobile devices.</em></td>
<td>Single User Driven</td>
</tr>
<tr>
<td><a href="http://webdesignerwall.com/">URL: http://webdesignerwall.com/</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technology Education Know-How</td>
<td>A blog to help teachers use technology. Maintained by a single author who tags the blog entries. Task: <em>Find two posts to assist your friend (4th grade teacher) with resources to supplement his teaching.</em></td>
<td>Single User Driven</td>
</tr>
</tbody>
</table>

When participants were asked which tool retrieved the most relevant results; 43% selected search, 43% selected tag clouds and only 14% selected menus. 50% of the participants indicated search to be the easiest navigation tool to use, 43% preferred using it and 29% considering them efficient in finding information, as shown in Figure 4.

**Fig. 4: Participant’s Perception of the Three Tools**

<table>
<thead>
<tr>
<th></th>
<th>Search</th>
<th>Menu</th>
<th>Tag Cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easiest</td>
<td>50%</td>
<td>14.29%</td>
<td>35.71%</td>
</tr>
<tr>
<td>Preferred</td>
<td>42.86%</td>
<td>28.57%</td>
<td>28.57%</td>
</tr>
<tr>
<td>Efficient</td>
<td>28.57%</td>
<td>28.57%</td>
<td>42.86%</td>
</tr>
</tbody>
</table>
Overall tag clouds were found to be comparable to search and 36% of the participants indicated it was the easiest navigation tool to use, 29% preferred using it and 43% considering them efficient in finding information. Menu scored the worst with only 15% considering them easy to use and only 29% finding them efficient. Participants especially disliked using menus when they considered them to be poorly labeled or when they are not complete enough for the task.

When participants were given a choice to use any of the three navigation tools, search by itself was used 32% of the time, 21% of the time in combination with menu and 11% of the time in combination with tag cloud. Menus and tag clouds by itself were used 11% of the time and 7% of the time in combination with each other. All three navigation tools were used 7% of the time, as shown in Figure 5.

The One-way Analysis of Variance (ANOVA) test was used to compare the two types of tag clouds. It was determined that single user driven tag clouds are efficient in terms of the time to complete the task than community driven tag clouds at a significance level of 0.075. Moreover, single user driven tag clouds required fewer clicks than community driven tag clouds at a significance level of 0.002.

Navigation on websites should, of course, be both intuitive and effective in locating relevant information. The findings from this and other studies, both qualitative and quantitative, reveal that users perceive tags and tag clouds to be as useful as search in finding information. The study provides sufficient evidence to continue to investigate the role of social tagging in the local classification of web content, especially for those websites that rely on community driven content.

Based on the findings from the study, we make the following suggestions for future work. The results suggest that an integration of the two navigation tools (tag cloud and search) might allow users the ability to search within a tag or to select multiple tags within a tag cloud, similar to multi-term search functions. The frequency of similar variations of tags could be mitigated by allowing users to select a tag from a list of recommended tags rather than always creating a new tag. Finally, a tool is suggested that guides navigation choices on recommendations from friends in the user's social network to provide more direct access to items of interest.

Challenges remain especially related to the integration or complementarity of the rich and robust established classification schemes with the more spontaneous social classification approach. For example, tags may be generated by users based on underlying structures, such as faceted schemes or hierarchies. The reality of user generated tags does raise important challenges for both retrieval and classification of the information that will see new applications of conventional classification to social classification.
References
M. Cristina Pattuelli - Pratt Institute, New York, NY USA
Sara Rubinow - Pratt Institute, New York, NY USA

Charting DBpedia:
Towards a Cartography of a Major Linked Dataset

Abstract
This paper provides an analysis of the knowledge structure underlying DBpedia, one of the largest and most heavily used datasets in the current Linked Data landscape. The study reveals an evolving knowledge representation environment where different descriptive and classification approaches are employed concurrently. This analysis opens up a new area of research to which the knowledge organization community can make a significant contribution.

Introduction
Linked Data extends the traditional web by providing an open and unifying framework for the discovery, integration, and reuse of data. It has the potential to realize the vision of the Semantic Web by promoting interoperability through the interlinking of well-defined machine-readable data. One of the strengths of the Linked Data initiative lies in its technical infrastructure, which is based on a simple set of principles and open standards. These standards include the Uniform Resource Identifier (URI), which serves as the global identifier mechanism, and the Resource Description Framework (RDF), which acts as the model for expressing data in a common format (Berners-Lee, 2009). This lightweight framework is key to lowering the barriers to Semantic Web adoption.

The library and cultural heritage communities are actively participating in Linked Data research and development, particularly in the context of its open license version, Linked Open Data (LOD). LOD is seen as a promising technology for connecting data and metadata from heterogeneous sources and contexts in order to facilitate their integration and sharing (Baker et al., 2011).

Major efforts of the library community in RDF development are currently focused on converting bibliographic data, including authority vocabularies, into RDF triples, the building blocks of Linked Data. The purpose of these efforts is to steer library data into a seamless search stream. Becoming part of a single global data space would be a significant departure from the traditional “walled garden” model of the bibliographic universe. Library, archive and museum data would be interlinked with billions of RDF triples from a variety of external datasets, most of which rely on relaxed knowledge structures and approximate semantics.

This new and boundless scenario necessitates a strategic reconsideration of both the nature of knowledge organization systems and the role they play in RDF services. In this evolving information environment, the notion of the knowledge structure is increasingly dynamic and flexible, though often less coherent and consistent. The RDF initiative, especially within the library and cultural heritage communities, is still in its infancy. We currently lack an in-depth understanding of the nature and function of semantics and knowledge structures in the new context of RDF application. Interlinking sound datasets with less formalized RDF datasets requires new ways of defining and evaluating the functional requirements and data quality standards of knowledge organization systems.

The literature on Linked Data development has just begun to address the implications of dealing with loosely formalized knowledge structures that produce significant amounts of “noisy” data. For example, issues related to the proliferation of identifiers assigned to the
same entity, as well as to the inappropriate use of the OWL\(^1\) identity property when managing co-references, have been discussed (Halpin et al., 2010; Hayes, 2011; Uschold, 2010).

In order to develop an understanding of the nature of the new knowledge systems emerging in the context of the RDF landscape, this paper investigates the semantic structure underlying DBpedia\(^2\), one of the largest and most heavily used RDF datasets. We focused specifically on the domain of jazz within DBpedia, which served as the application scenario from which to begin to develop a “cartography” of this semantic area. Jazz was chosen due to the researchers’ familiarity with the subject area in the context of RDF technology applications.

**DBpedia**

DBpedia is a multilingual and cross-domain RDF dataset created with the purpose of making Wikipedia information available as Linked Data. Currently, the English version of DBpedia includes over 3.64 million resources, each referring to an article of Wikipedia. DBpedia has become one of the most important repositories of Linked Data semantics. It is interlinked to other datasets and heavily used as a semantic hub, as illustrated by the Linking Open Data cloud diagram (Cyganiak & Jentzsch, 2011).

The main source of structured data in Wikipedia is the infobox, which is located in the upper right-hand corner of a Wikipedia article. The infobox presents relevant information about a Wikipedia article in a standardized format. Infoboxes are key to understanding both the origin of the properties that populate the DBpedia knowledge base and to gaining a better sense of the structure underlying DBpedia.

Infoboxes are based on a wide range of templates that contain core information elements related to the subject of Wikipedia articles. More than 6,000 templates exist within English-language Wikipedia\(^3\), and that number is constantly changing. Infobox templates are created and reused by Wikipedia contributors who also supply the documentation and the rules that determine their format and use. The lack of centralized administrative control has implications for the consistency and accuracy of how the information fields within infobox templates are filled as well as how those templates are utilized. As Bizer et al. (2009) note, Wikipedia authors may use different templates to describe the same topic; the same attribute can be named differently in different templates (e.g., birthPlace and placeOfBirth); and attribute values may be expressed using different formats.

**DBpedia Data Extraction**

Two extraction methods, generic and mapping-based, are employed to acquire structured information from Wikipedia templates (Bizer et al., 2009). The generic method is algorithm-based and consists of bootstrapping the complete coverage of the infobox content and retaining the same property names used in the infobox. The data obtained through this process are described by properties identified by the dbpprop prefix. Because the dbpprop properties are not part of any cohesive structure or conceptual model, they bring to the dataset a good deal of redundancy and inconsistency.

In order to increase precision and consistency, an additional extraction process was later adopted that is based on manually generating mappings between Wikipedia infobox templates and the DBpedia Ontology\(^4\). This method was introduced in an effort to contribute a set of controlled, higher quality data to complement the raw, noisy dataset.

Like Wikipedia, DBpedia grants editing rights to anyone motivated to create manual mappings of Wikipedia infobox templates. The majority of Wikipedia data remains to be

---

\(^1\) [http://www.w3.org/TR/owl-features/](http://www.w3.org/TR/owl-features/)

\(^2\) [http://dbpedia.org/](http://dbpedia.org/)


\(^4\) [http://wiki.dbpedia.org/Datasets](http://wiki.dbpedia.org/Datasets)
mapped, and crowdsourcing is a solution that has been adopted to increase efficiency by distributing the work across subjects and languages. Soon after editing rights were opened to the community at large in the summer of 2011, the DBpedia team noted the proliferation of redundant mapping, expressing concerns that “the ontology is getting unclear and the benefits of standardisation get lost.” Achieving a balance between mapping coverage and ontology clarity appears to be a goal that remains unresolved.

The DBpedia Ontology

The DBpedia Ontology is a general ontology that covers multiple domains. It consists of a shallow class hierarchy of 320 classes created by manually deriving 170 classes from the most popular Wikipedia’s infobox templates in the English edition. It also includes 750 properties resulting from mapping attributes from within these templates. The DBpedia Ontology currently describes approximately half of the DBpedia entities (DBpedia, 2011).

Because the DBpedia Ontology is built upon infobox templates, its classes suffer from a lack of logical consistency and present significant semantic gaps in their hierarchy. As Damova et al. (2010) point out, inconsistencies are shown, for example, in the degree of generality present in its upper-level classes that range from abstract concepts such as person or event to very specific ones like protein or drug.

The concept jazz is not included in the ontology. The broader domain of music is currently represented by eight classes: MusicalArtist, MusicFestival, MusicGenre, Musical, MusicWork, Album, Single, Song. Each class is part of an unrelated segment of the tree structure that has owl: Thing as its root. Table 1 shows a sample of the knowledge structure of the ontology.

Table 1: A sample of classes and properties in the domain music

<table>
<thead>
<tr>
<th>DBpedia Ontology Classes</th>
<th>DBpedia Ontology Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>● owl:Thing</td>
<td></td>
</tr>
<tr>
<td>○ Person</td>
<td>almaMater, knownFor, occupation, residence</td>
</tr>
<tr>
<td>● Artist</td>
<td>nationality, influencedBy, award, training</td>
</tr>
<tr>
<td>● MusicalArtist</td>
<td>birthDate, birthPlace, hometown, instrument</td>
</tr>
<tr>
<td>● owl:Thing</td>
<td></td>
</tr>
<tr>
<td>○ MusicGenre</td>
<td>stylisticOrigin, musicSubgenre, musicFusionGenre</td>
</tr>
</tbody>
</table>

DBpedia data represented by the DBpedia Ontology are identified by the prefix dbpedia-owl. As with the uncontrolled dbpprop properties, a comprehensive dictionary of dbpedia-owl properties is not readily available. A full set of infobox-based dbpedia-owl properties can be viewed on an infobox-by-infobox basis through the graphical user interface of the DBpedia Mapping Tool.

To attempt to understand how DBpedia properties, both ontology-based and uncontrolled, operate in the domain of jazz, we focused on the description of jazz musicians. In the absence of a general inventory of DBpedia properties, we loaded the musical artist infobox template in the DBpedia Mapping Tool in order to discover the full set of dbpedia-owl properties employed to describe jazz artists.

---

6 http://mappings.dbpedia.org/index.php/Mapping_Guide
7 http://mappings.dbpedia.org/index.php/MappingTool
Table 2 shows the list of dbpedia-owl properties derived from the ontology-based mapping of the musical artist infobox template. These properties are presented alongside the corresponding dbpprop properties. The names of the dbpprop properties mirror the original name of the corresponding Wikipedia infobox template field.9

Table 2: Properties describing a MusicalArtist entity through the musical artist infobox template

<table>
<thead>
<tr>
<th>Properties from mapping-based extraction</th>
<th>Properties from generic extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbpedia-owl:activeYearsEndYear</td>
<td>dbpprop:years_active</td>
</tr>
<tr>
<td>dbpedia-owl:activeYearsStartYear</td>
<td>dbpprop:years_active</td>
</tr>
<tr>
<td>dbpedia-owl:alias</td>
<td>dbpprop:alias</td>
</tr>
<tr>
<td>dbpedia-owl:associatedBand</td>
<td>dbpprop:associated Acts</td>
</tr>
<tr>
<td>dbpedia-owl:associatedMusicalArtist</td>
<td>dbpprop:associated Acts</td>
</tr>
<tr>
<td>dbpedia-owl:background</td>
<td>dbpprop:background</td>
</tr>
<tr>
<td>dbpedia-owl:birthDate</td>
<td>dbpprop:birth_date</td>
</tr>
<tr>
<td>dbpedia-owl:birthPlace</td>
<td>dbpprop:birth_place</td>
</tr>
<tr>
<td>dbpedia-owl:deathDate</td>
<td>dbpprop:death_date</td>
</tr>
<tr>
<td>dbpedia-owl:deathPlace</td>
<td>dbpprop:death_place</td>
</tr>
<tr>
<td>dbpedia-owl:genre</td>
<td>dbpprop:genre</td>
</tr>
<tr>
<td>dbpedia-owl:hometown</td>
<td>dbpprop:origin</td>
</tr>
<tr>
<td>dbpedia-owl:instrument</td>
<td>dbpprop:instrument</td>
</tr>
<tr>
<td>dbpedia-owl:occupation</td>
<td>dbpprop:occupation</td>
</tr>
<tr>
<td>dbpedia-owl:recordLabel</td>
<td>dbpprop:label</td>
</tr>
</tbody>
</table>

As discussed earlier, the concept jazz is not part of the DBpedia Ontology. Jazz is the subject of a Wikipedia article and, as such, it is an entity in the DBpedia knowledge base. The Wikipedia entry for jazz includes a music genre infobox, from which DBpedia can extract structured data. As with all DBpedia entities, the DBpedia properties for jazz are displayed on a corresponding HTML DBpedia resource webpage, which lists data representing the entity in a human-friendly fashion. The data are presented in the form of property-value pairs.

It would be expected that the DBpedia resource page for jazz would include all the dbpedia-owl and dbpprop properties associated with any other MusicGenre entity (e.g., instrument, stylisticOrigin). Upon reviewing the jazz resource page, however, none of these properties were found. The properties are present, however, on other DBpedia resource pages for entities that also employ the music genre infobox template, such as Bebop, Swing music and Rhythm and blues. This discrepancy does not seem to stem from the data source, since the infobox fields in the Wikipedia entry for jazz are filled appropriately. It is more likely that the properties were missed during the extraction process, which is not an uncommon occurrence.

Knowledge Representation Tools in DBpedia

In addition to the DBpedia Ontology, the DBpedia knowledge base is governed by a variety of knowledge representation tools including additional classification schemes and RDF vocabularies.

---

9 It should be noted that the following three dbpedia-owl properties are not based on infobox template mappings and thus not included in the table: dbpedia-owl:abstract, dbpedia-owl:thumbnail and dbpedia-owl:wikiExternalLinks.

10 http://dbpedia.org/page/Jazz
Two classification systems are particularly relevant and consistently used: Wikipedia categories and the YAGO ontology. Wikipedia uses categories to group articles that share similar subjects. Wikipedia categories are constantly evolving and currently number more than 740,000. Most categories are assigned manually by Wikipedia contributors and can be found listed as links at the bottom of a Wikipedia article. For example, the page about the jazz artist Mary Lou Williams currently displays 28 categories, ranging from Musicians from Pittsburgh and Converts to Roman Catholicism to Deaths from bladder cancer. Each category links to a category page containing an alphabetized list of links to other Wikipedia articles assigned to the same category. When available, the category page also lists related parent- and sub-categories.

Wikipedia categories are organized hierarchically, but they are not grounded in a strict taxonomic structure. For example, “Mary Lou Williams” is assigned simultaneously to three categories related to music composition, all of which should be in a subset relation: American composers, Jazz composers, and Women composers. As Suchanek, Kasneci and Weikum (2008) point out, this category system merely mirrors the “thematic structure” of a Wikipedia article rather than representing a cohesive knowledge conceptualization (p. 210).

The Wikipedia organizational system is the result of a collaborative effort that presents advantages as well as weaknesses. On one side, the Wikipedia authoring and editorial process ensures that the categories are continually updated to correspond with article content. On the other side, the system suffers from lack of consistency in its hierarchical structure and what Bizer et al. (2009) call a “rather loose relatedness between articles” (p. 157).

DBpedia makes use of Wikipedia categories to organize its entities. The hierarchical structure of the categories is represented in DBpedia by way of two different properties: dcterms:subject and skos:broader. The property dcterms:subject relates DBpedia entities to their corresponding categories. Each category is then related to its parent category through the skos:broader property (Mirizzi, Di Noia, Ostuni, & Ragone, 2012).

The YAGO ontology is another classification system introduced to provide DBpedia data with coherence and structural consistency. YAGO, the most recent iteration of which is called YAGO2, was developed at the Max Planck Institute for Informatics in Saarbrücken, Germany. YAGO was originally derived from the Wikipedia category system using the semantic lexicon WordNet. More specifically, YAGO was created through the automatically generated mapping of Wikipedia categories to WordNet synsets (Suchanek, Kasneci & Weikum, 2007). It is a robust and extremely rich classification scheme with a deep hierarchical structure. As Bizer et al. (2009) note, while YAGO is very accurate, it is not immune to the errors and omissions that inevitably occur when ontologies are created using algorithm-based methods. DBpedia uses “YAGO as a taxonomic backbone to connect the facts to a coherent whole” (Suchanek et al., 2008, p. 2).

YAGO class instances are represented as values of the rdf:type property. The rdf:type property is also paired with values that are class instances of various external ontologies including OWL, schema.org and UMBEL. The ontology classes serve as connectors that facilitate the interlinking of web content and data to give context to these data.

Figure 1 shows clusters of DBpedia properties as well as external vocabularies that represent the DBpedia entity “Mary Lou Williams.”

---

12 http://wordnet.princeton.edu/
13 http://schema.org/
14 http://umbel.org/
Fig. 1: A sample of properties and vocabularies used to represent the DBpedia entity “Mary Lou Williams.”

Both RDF native and RDF-based vocabularies are also employed to describe DBpedia data, such as the Dublin Core Metadata Element Set (e.g., dc:description) and the Friend of a Friend (FOAF) ontology (e.g., foaf:name, foaf:givenName, foaf:page). The use of properties from different external vocabularies is made possible by the unifying framework provided by RDF. The flexibility of the RDF model allows for the mixing and matching of properties from different namespaces without the need for agreement on the adoption of a specific schema. Multiple vocabularies can be used in a layered fashion, and properties with overlapping scope can coexist in the spirit of “cooperation without coordination” (Wood, 2011), challenging the classical notion of semantic parsimony.
Conclusion
This paper investigates the semantic structure underlying DBpedia, one of the largest datasets in the context of RDF. Our analysis reveals a new type of knowledge representation environment that is in a constant state of flux, where different descriptive and classification approaches are employed concurrently.

The vocabularies and schemes employed in this open and dynamic environment vary significantly in terms of their degree of formalization, stability, cohesiveness and consistency. As such, they challenge our tolerance threshold for data quality and our traditional notion of authority control. Unearthing the knowledge organization of DBpedia increases our practical understanding of the new semantic context provided by RDF. It also has the potential to enhance the efficacy of publishers, application developers and vocabulary maintainers who work with Linked Data. For example, by clearly delineating the available range of properties, developers could leverage DBpedia semantics to formulate queries with higher precision, rather than use a trial and error approach.

This analysis of DBpedia opens up a new area of research to which the knowledge organization community can make a significant contribution. The decentralized interplay of vocabularies as well as the proliferation of noisy data have implications that have yet to be fully understood. As Dunsire et al. (2011) stress, there will be an increasing need to understand and leverage what is perceived as chaos, rather than fearing the presumed end of an existing order.

References


Subject Organization in Three Types of Information Resources: An Exploratory Study

Abstract
Knowledge tends to be structured differently in different types of information resources and information genres due to the different purposes of the resource/genre, and the characteristics of the media or format of the resource. This study investigates subject organization in three types of information resources: books (i.e. monographs), Web directories and information websites that provide information on particular subjects. Twelve subjects (topics) were selected in the areas of science, arts/humanities and social science, and two books, two Web directories and two information websites were sampled for each subject. The top two levels of the hierarchical subject organization in each resource were harvested and analyzed. Books have the highest proportion of general subject categories (e.g. history, theory and definition) and process categories (indicating step-by-step instructions). Information websites have the highest proportion of target user categories and genre-specific categories (e.g. about us and contact us), whereas Web directories have the highest proportion of specialty categories (i.e. sub-disciplines), industry-role categories (e.g. stores, schools and associations) and format categories (e.g. books, blogs and videos). Some disciplinary differences were also identified.

Introduction
In an information resource (e.g. a book or website), information is organized to serve the purpose of the resource, to make the content logically coherent and to help the reader/user to access and assimilate the information effectively. The overall information organization is usually hierarchical, and may be presented as a table of contents (in the case of a book), a menu system (for a Website), or a hierarchically organized set of hyperlinks (for a Web directory or the site map of a website). These give the reader/user an overview of the information content of the resource - especially the main topics and sub-topics and their relationships. They also support navigation and browsing, allowing the user to quickly zoom into a particular subtopic of interest.

Information on a particular subject is likely to be organized differently in different types of information resources and information genres to serve different purposes and different target users. The characteristics and constraints of the format or media may also impact the organization of information. The purpose of the study is to find out how information or knowledge tends to be organized in three types of information resources:

1. books on a particular subject, represented by their table of contents
2. information websites on a subject, represented by their menu systems
3. Web directories of resources on the Web, represented by their hierarchically organized hyperlinks to resources on a subject.

These resources have either a pedagogical purpose to impart knowledge about a particular subject, or a reference function to link users to a range of resources related to the subject. Thus information organization in these resources can be considered as knowledge organization of a particular subject or field, from a particular perspective. The knowledge
organization is assumed to be hierarchical and can be represented by a taxonomy (or tree structure).

A book (monograph) on a particular subject is limited to a few hundred pages and tends to have a pedagogical purpose to impart knowledge about a subject. Readers are expected to read the book from beginning to the end in sequence, or at least a substantial portion of the content (e.g. a chapter). A book is organized into chapters and chapters are subdivided into sections.

An information website provides information on a subject, and may serve a combination of pedagogical and reference functions. It is not limited by the number of pages and moreover has hyperlinks that link to other Web resources. Information is presented in smaller self-contained chunks that fit on a webpage that is two or three-screens in length. The reader is not expected to read all the pages or read them in a particular sequence.

A Web directory serves primarily a reference function. It does not have much native content stored on the resource, but has a large number of hyperlinks that refer the user to many different types of resources on the Web. It is likely to have a large knowledge structure because providing a knowledge structure is its main function.

In analyzing the information organization in the three types of resources, we attempted to identify recurring patterns and facets (category types) that occur across resources and differences between the types of resources. As the information organization may be affected by the subject area, we selected 12 subjects of different types - two broad subjects and two narrower subjects each from the sciences, arts/humanities and social sciences. Two books, two Web directories and two information websites were sampled for each subject. To keep the analysis manageable, only the top two levels of the knowledge structures (taxonomies) were analyzed. Kwasnik (1999) referred to this as the “first cut” that determines the overall shape and representational eloquence of the classification.”

This is part of a series of studies we are carrying out on knowledge organization in different environments and for different purposes.

Previous Studies

There are surprisingly few studies of information/knowledge organization in various information genres and media. Most studies of websites have focused on usability and interface design. Hardly any have focused on the navigation structure, other than to determine the importance of ease of navigation (e.g., Parboteeah, Valacich & Wells, 2009; Tarafdar & Zhang, 2005) and to present general guidelines for the design and presentation of the navigation structure (e.g., Oppenheim & Ward, 2006; Perugini & Ramakrishnan, 2006).

Chen, Magoulas and Dimakopoulos (2005) studied users’ cognitive styles and their preference for different kinds of hierarchical structures for browsing Web directories. They compared two cognitive styles: field independence and field dependence. Based on their study of three Web directories - Google, Alta Vista and Lycos - they concluded that field dependent users, who typically see the global picture and ignore details, prefer a wide and shallow hierarchical structure (with many main categories and few levels of subcategories), and prefer the main categories and subcategories to be presented on different screens. On the other hand, field independent users, who are less dependent on the context and tend to focus on details, prefer a narrow and deep hierarchy (with a few main categories and many levels of subcategories), and prefer the main categories and subcategories to be displayed on the same screen.

Kwasnik (2002) analyzed the classification scheme used in Amazon.com to organize books. She found a rich network of terms in an enumerative and multihierarchical structure” and multiple paths to the same books. At the top level, topical categories were
mixed with genre categories, and categories reflecting perspectives, formats and promotions/offers. There was no consistency in scale, granularity, terminology or structure. The structure was pragmatic - to increase the likelihood of a user making a purchase - rather than to provide a knowledge organization structure. Chang (2004) analyzed the types of category redundancies found in the classification schemes of Amazon.com and the U.S. Energy Information Administration website. She found that over half of the categories were redundant, and identified five types of redundancies that suggest different kinds of issues. The redundant categories support the pragmatic purposes of the websites by providing multiple entry points and paths to the same item.

In this study, the books, Web directories and information websites that were analyzed seek mainly to provide information, though some of the information is of services, organizations and personalities. We argue that the information organization structure in these resources reflect the author/designer’s knowledge structure, which includes a high-level model of the academic discipline (the first cut) as well as industry knowledge. The extent that this coincides with the user/reader’s conception probably varies. But for the information organization structure to be useful, it should be compatible with the user’s conception, or at least something the user can grasp. However, we concede that this issue needs further analysis and discussion, and we hope that the results of our study will contribute to this discussion.

Research Method
The selection of the subjects started with the list of academic disciplines in Wikipedia (http://en.wikipedia.org/wiki/List_of_academic_disciplines). Six broad subjects and six narrower subjects were selected, based on the authors’ interests:

1. From the Health/Medicine discipline:
   • broad subjects: health and medicine
   • narrower subjects: cancer and surgery

2. Music discipline:
   • broad subjects: string instruments and folk music
   • narrower subjects: violin and flamenco music

3. Education discipline:
   • broad subject: education
   • narrower subject: child education.

4. Psychology discipline:
   • broad subject: psychology
   • narrower subject: child psychology.

For each of the subjects, we selected two books, two Web directories and two information websites. Altogether, 72 (12x6) resources were analyzed. The resources were identified by carrying out title-keyword searches in a library catalog and searching for the subject keyword in Google.

From each of the resources, we harvested level 1 and level 2 category items from the knowledge structure of the resource: the table of contents for books, the hierarchical menu structure from information websites, and the hierarchy of hyperlinks from Web directories.

From a preliminary analysis of the knowledge structures, we identified seven recurring facets in the organization of information:

1. Specialty facet
2. General subject facet
3. Industry role facet
4. Process facet
5. Target user facet
6. Format facet

These are explained in the next section. They are used as a framework to analyze the similarities and differences in information organization in books, Websites and Web directories.

Results
Table 1 gives the distribution (proportion) of categories from the seven facets - for books, Web directories and information websites. Table 2 provides more detail, listing the average number of categories of each facet that can be expected in each type of resource. Of the three types of information resources, Web directories have the biggest breadth in knowledge structure with an average of 15.3 categories at level 1 and 31.0 at level 2 (see Table 2). Books are next with an average 12.9 chapters and 22.6 chapter sections. Information websites have an average of 9.0 categories at level 1 and 23.1 categories at level 2. The three types of information resources have different profiles in the distribution of categories from the seven facets. We discuss each facet in more detail.

Table 1: Percentage of categories from the seven facets for the three types of resources

<table>
<thead>
<tr>
<th>Facet</th>
<th>Book</th>
<th>Information website</th>
<th>Web directory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specialty facet</td>
<td>46.4%</td>
<td>40.7%</td>
<td>60.7%</td>
</tr>
<tr>
<td>General subject facet</td>
<td>37.1%</td>
<td>30.7%</td>
<td>14.4%</td>
</tr>
<tr>
<td>Industry role facet</td>
<td>2.5%</td>
<td>5.3%</td>
<td>12.2%</td>
</tr>
<tr>
<td>Process facet</td>
<td>2.9%</td>
<td>0.9%</td>
<td>0.7%</td>
</tr>
<tr>
<td>Target user facet</td>
<td>1.5%</td>
<td>8.1%</td>
<td>4.7%</td>
</tr>
<tr>
<td>Format facet</td>
<td>0</td>
<td>2.7%</td>
<td>3.9%</td>
</tr>
<tr>
<td>Genre-specific category</td>
<td>9.6%</td>
<td>11.5%</td>
<td>3.4%</td>
</tr>
</tbody>
</table>

Total 100% 100% 100%

* The highest percentage in each row are highlighted in italics

Table 2: Average number of categories of each facet in an information resource, for the 3 types of resources

<table>
<thead>
<tr>
<th>Facet</th>
<th>Specialty</th>
<th>General subject</th>
<th>Industry role</th>
<th>Process</th>
<th>Target user</th>
<th>Format</th>
<th>Genre-specific</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book Level 1</td>
<td>5.58</td>
<td>4.42</td>
<td>0.08</td>
<td>0.29</td>
<td>0.16</td>
<td>0</td>
<td>2.33</td>
<td>12.86</td>
</tr>
<tr>
<td>Level 2</td>
<td>10.87</td>
<td>8.75</td>
<td>0.79</td>
<td>0.75</td>
<td>0.37</td>
<td>0</td>
<td>1.08</td>
<td>22.61</td>
</tr>
<tr>
<td>Total</td>
<td>16.46</td>
<td>13.17</td>
<td>0.87</td>
<td>1.04</td>
<td>0.54</td>
<td>0</td>
<td>3.42</td>
<td>35.50</td>
</tr>
<tr>
<td>Information website Level 1</td>
<td>2.96</td>
<td>2.08</td>
<td>0.58</td>
<td>0.08</td>
<td>1.17</td>
<td>0.46</td>
<td>1.71</td>
<td>9.04</td>
</tr>
<tr>
<td>Level 2</td>
<td>10.13</td>
<td>7.79</td>
<td>1.13</td>
<td>0.21</td>
<td>1.46</td>
<td>0.42</td>
<td>2</td>
<td>23.14</td>
</tr>
<tr>
<td>Total</td>
<td>13.08</td>
<td>9.88</td>
<td>1.71</td>
<td>2.63</td>
<td>1.88</td>
<td>3.71</td>
<td>3.71</td>
<td>32.17</td>
</tr>
<tr>
<td>Web directory  Level 1</td>
<td>10.54</td>
<td>2.25</td>
<td>1.08</td>
<td>0</td>
<td>0.46</td>
<td>0.50</td>
<td>0.42</td>
<td>15.25</td>
</tr>
<tr>
<td>Level 2</td>
<td>17.54</td>
<td>4.42</td>
<td>4.58</td>
<td>0.33</td>
<td>1.71</td>
<td>1.29</td>
<td>1.17</td>
<td>31.04</td>
</tr>
<tr>
<td>Total</td>
<td>28.08</td>
<td>6.67</td>
<td>5.67</td>
<td>0.33</td>
<td>2.17</td>
<td>1.79</td>
<td>1.58</td>
<td>46.29</td>
</tr>
</tbody>
</table>

Note: The figure in each cell is an average from 24 resources
Specialty Facet
The specialty facet refers to subdivision into sub-disciplines. Sub-disciplines are domain dependent—they are different from one subject to another. For example, medicine was subdivided into cancer, heart disease, surgery, etc. and music was subdivided into instruments, genres, etc.

Table 1 shows that specialty categories are the most common for all three resource types, accounting for 40% to 60% of the categories. Web directories have the highest proportion of categories by specialty, accounting for 60.1% of the categories in level 1 and 2. For example, the Web directory —Education Planet” has nine level 1 categories, eight of which are sub-disciplines such as math, art and music, and geography. This is probably because Web directories seek to be fairly comprehensive in its coverage of relevant resources on the Internet, and carries minimal native content on its site.

General Subject Facet
The general subject facet refers to common subject subdivisions such as history, definition and education that can be used to subdivide any subject. General subject is the next most frequent type of category after specialty, often found in books (37%) and websites (30%), and less frequently in Web directories.

These common subjects, e.g. history and definition, serve a pedagogical function and provide a foundation for understanding a subject. They are thus more likely to be found in books and information websites. For example, the book —Introduction to Education Studies” has chapters on —What is education studies?” (i.e. definition), —The nature of education” and —Modern history of schooling” at level 1. The book —Education Studies: An Introduction” has chapters on —Theoretical perspectives” and —Policy.” On the other hand, Web directories serve primarily a reference function and have fewer categories devoted to introductory or tutorial material.

Industry Role
The industry role categories are practical in nature and refer the reader to available services. In the medicine domain, it includes hospitals, clinics and medical associations; the music domain has music stores, music schools and luthiers; and the education domain has kindergartens and internships. Industry role categories are mainly found in Web directories (12.2% of categories), and to a smaller extent in information websites (5.3%).

Process Facet
Process categories indicate content that provide step-by-step instructions. Examples in the medicine domain are —7 steps for cancer prevention” and —How to lose weight”; and the music domain has —Learn to play violin,” —How to make a violin” and —Suzuki violin method.” Process categories occur most often in books (2.9% of categories) as the categories indicate a teaching function.

Target User Facet
Target user categories indicate information that is oriented to a particular target audience. For example, the medicine domain has categories —For doctor” and —For patient”; the music domain has categories —For child” and —For adult”; and the education domain has categories —For teacher” and —For student.” These categories occur most often in information website (8.2%), followed by Web directories (4.7%).

Format Facet
Format includes categories such as books, blogs, videos, albums and Flash. These categories are more often found in Web directories (3.9%) and websites (2.7%), but seldom in books.
Genre-specific categories

Genre-specific categories are categories that are associated with particular media or genre. For example, books often have *preface*, *bibliography* and *index*; information websites often have *home*, *about us*, *contact us* and *site map*; and Web directories often have *links* and *personal pages*. These genre-specific categories are more often found in books (9.6%) and websites (11.53%), and less often in Web directories (3.4%).

Differences Across Disciplines

We compared the proportions of the facets across the disciplines and found some interesting differences. The differences may be random effects due to the small sample size. However, they suggest issues to explore further.

We found the music discipline to have a higher proportion of *specialty* categories than expected and lower proportion of *general subject* categories:

- For music information websites: a high 65% of specialty categories (compared to 41% for all 4 disciplines) and a low 9% of general subject categories (compared to 31% for all 4 disciplines)
- For music books: a high 66% of specialty categories (compared to 46% overall) and a low 25% of general subject categories (compared to 37% overall).

In contrast, education and psychology have a lower proportion of *specialty* categories and higher proportion of *general subject* categories:

- For education/psychology information websites: a low 28% of specialty categories and a high 39% of general subject categories
- For education/psychology books: a low 39% of specialty categories and a high 51% of general subject categories.

This suggests that the music resources sampled are more specialized or address a more specialized audience, whereas the education and psychology resources address the general public. In fact, the education and psychology websites in our sample carry a substantial amount of educational resources for parents and teachers. This may reflect the nature of the disciplines or may be an artifact of the sampling.

The medicine and music Web directories in our sample contain a high 20% of *industry role* categories (with links to organizations, products and services), whereas the education and psychology Web directories contain a low 1% of *industry role* categories (with more links to educational resources).

Conclusion

This exploratory study has sampled only a small number of disciplines and a small number of resources. Though it is not possible to make statistical generalizations, the results are suggestive and do make sense. Books, Web directories and information websites have different information organization structures, according to the different purposes they serve. Books have a higher proportion of *general subject* categories and *process* categories as they serve a pedagogical purpose of imparting basic knowledge on a subject. They also have genre-specific categories such as *preface*, *bibliography*, *appendix* and *index*.

Information websites have a higher proportion of *target user* categories. As they are easily accessible on the Internet and are visited by users with different needs and levels of expertise, they may attempt to serve both experts (e.g. doctors) and lay users (e.g. patients). Websites also have a high proportion of web-specific categories such as *about us*, *contact us* and *site map*.

Web directories have the highest proportion of specialty categories (i.e. sub-disciplines) as they serve a reference function and attempt to provide an extensive coverage of relevant
resources on the Internet. They also serve practical functions with industry-role categories that help users to locate services, and have more format categories indicating coverage of resources of different media and formats.

These results need to be confirmed by a larger-scale systematic study. Disciplinary differences are likely to be found, and indeed some disciplinary differences were found in our sample.

The facets adopted in this study were derived rather informally from a preliminary review of the categories harvested from the sample resources. The facets appeared salient probably because of the researchers’ familiarity with the purposes and characteristics of the different information genres. The seven facets were found to be useful in distinguishing between the types of resources, but there are conceivably other ways of grouping categories into facets and other potentially useful facets. Focus groups can be carried out to identify other facets to investigate. In addition, frequency analysis can used to identify common categories which may reflect facets. Some facets are likely to be discipline dependent. Facets that are important in the music domain (e.g. genre, instrument and time period) may not be important in other domains. In this study, the categories were assigned to facets manually but no inter-coder reliability test was carried out. Making a list of categories under each facet will help to improve coding reliability.

As the specialty facet accounts for the highest proportion of categories, future work can examine this facet to identify differences in the way the different resources divide a subject into sub-disciplines. Some of the categories coded in this study as specialty categories can be considered to be separate facets, e.g. instrument and genre in the music discipline. It will also be revealing to compare the subject organization found in the different information genres with the knowledge structures in general classification schemes such as the Library of Congress Classification Scheme and Dewey Decimal Classification.

References
A Novel Knowledge Organization Scheme for the Web: Superlinks with Semantic Roles

Abstract
We discuss the needs of a knowledge organization scheme for supporting Web-based software applications. We show how it differs from traditional knowledge organization schemes due to the virtual, dynamic, ad-hoc, user-specific and application-specific nature of Web-based knowledge. The sheer size of Web resources also adds to the complexity of organizing knowledge on the Web. As such, a standard, global scheme such as a single ontology for classifying and organizing all Web-based content is unrealistic. There is nevertheless a strong and immediate need for effective knowledge organization schemes to improve the efficiency and effectiveness of Web-based applications. In this context, we propose a novel knowledge organization scheme wherein concepts in the ontology of a domain are semantically interlinked with specific pieces of Web-based content using a rich hyper-linking structure known as Superlinks with well-defined semantic roles. We illustrate how such a knowledge organization scheme improves the efficiency and effectiveness of a Web-based e-commerce retail store.

Introduction
The Web is a vast, highly dynamic, multilingual, multicultural social knowledge repository. It is forcing rapid changes in the ways by which knowledge is created, organized, shared and found. For the most part, the Web has grown both in size and in the variety of its uses without waiting for underlying theories, structures, standards and schemes of organization to catch up. Significant portions of the content on the Web are not well-edited or well-organized. Yet, its social and democratic nature coupled with omnipresent functionality such as search and navigation has created a vast level playing field for mankind to manage its knowledge in previously unthinkable ways. It is impractical, at this point of the growth of the Web, to propose a universal knowledge organization scheme for the entire Web.

In the present work, we focus on Web-based software applications such as electronic-commerce stores on the Web and investigate how knowledge is organized in those Web sites. Finding several deficiencies therein, we propose a novel knowledge organization (KO) scheme for such Web-based applications. The primary objective of a Web-based store (say for books or electronic gadgets) is to ensure that customers who visit their Web sites find (and purchase) what they are looking for. Additionally, it is important that customers do not place orders for incorrect or incompatible products given that they do not have the benefits of either examining the products physically or consulting a human salesperson before placing their orders. How should the store organize all of the knowledge about their products so that these requirements are met?

Currently, most Web-based stores are using ad-hoc taxonomies to classify their products. To overcome any inefficiency in navigating through the taxonomical categories, they provide a keyword search facility, and, sometimes even links to related categories and products. Links are standard hyperlinks, that is, clicking the link takes the user to a single approximate location on the Web. When a user has a more specific need, for example, to select a battery that is compatible with the particular model of a cell phone, even where such knowledge is available, it is usually embedded deeply in a document on the site that outlines all the specifications of the cell phone or the battery in unstructured content. The KO scheme is typically not rich enough to provide a direct link clicking on which takes the user to all compatible accessories for the product.
Additional reasons for requiring a good KO scheme for the Web include improving search quality, ease of browsing, enabling effective social knowledge management, better social networking and enabling newer semantic applications on the Web.

**Six Principles of Knowledge Organization for the Web**

In this section, we attempt to summarize our observations about the divergent needs of knowledge organization for traditional and Web-based applications in the form of six principles. It may be noted that the following principles are significantly different from those proposed earlier for organizing portals on the Web for information access by particular communities (Rowley, 2010).

**Principle 1: Virtuality:** *Knowledge Organization for the Web is virtual, not physical.*

In Web-based applications such as e-commerce, there is no physical organization of knowledge assets that could serve as a guiding structure for KO. At the same time, there are no constraints imposed by physical structure on KO. For example, the same item can be present simultaneously in multiple places in the virtual showcases of an e-commerce store which is impossible in a physical organization scheme.

Further, there is no constraint imposed by the structure of data or metadata. In traditional computerized applications, for example, the structure of data structures, tables and databases often constrains what information can be added. The inherently unstructured nature of Web-based applications with their flexible structures such as non-relational data stores and XML-based semi-structured schemas pose no such problems.

**Corollary:** *KO for the Web is inherently different from KO for primarily physical collections.*

In addition to virtuality, Web content is distributed, highly dynamic and large scale (also known as “Web-scale”). It often uses an uncontrolled vocabulary with no single acceptable hierarchy of well-defined concepts. The richness of meta-data and the resulting fine granularity of classification are also distinguishing features of KO schemes for the Web.

**Principle 2: Multiplicity:** *KO for the Web needs multiple classifications.*

Although a product may belong to only one category from an ontological point of view, single categories are not very effective in ensuring that customers can find the product easily. Nor is it possible to design a single set of facets that effectively classify all the products in a Web store.

The Web provides multiple ways of finding a resource. For example, if we perform a search on an e-commerce site like Amazon.com for a book titled “Diary of a Wimpy Kid,” we find that Amazon places this book under multiple categories such as *Children’s Books, Humour, Social Situations* and *Friendship.* Such a strategy of placing an item under multiple categories for better access is followed on almost all e-commerce sites. It is the “findability” of the book that matters to a user when it comes to retrieval.

**Principle 3: Ad-hoc-ness:** *Impure categories are necessary.*

Ontological purity in the form of set-theoretic subsumption relations among categories and sub-categories is less important than usability of the categories. For the needs of Web applications, we often have to create ad-hoc categories that are ill-defined (e.g., “hot-selling products,” “most-emailed stories,” etc.). Categories must be intuitive, natural, using common terms, readable, and easy to remember and recall. Standardization is an unrealistic goal in the rapidly changing world of Web commerce. What matters is how a category is going to be used in practice.
Principle 4: Customization: User-specific and localized categories are necessary.

Users need to be provided categories that are to their liking and are most appropriate to their specific local context. This is not the same as the community principle (Rowley, 2010). In an e-commerce application, for example, there may be no community, just individual customers who are often anonymous. In some situations, there may be extensive knowledge about customers from which one could recognize various demographic patterns. Continuing with the example of “Diary of a Wimpy Kid” on Amazon.com, hundreds of users have tagged the book under Books for Boys, Boys Books, So Funny, Comics, Cartoon, Comedy and Jeff Kinney (its author).

Principle 5: Ontology: KO for the Web needs ontological structure.

E-commerce on the Web requires organization of complex relationships and constraints among products. For example, cell phones, their hardware features, accessories and software applications have complex relationships of compatibility. Such interdependencies not only need to be represented in well-structured units in an ontology, they also need to be presented to the user through effective linking structures so that the user is guided efficiently in selecting the most appropriate products.

In addition, manual classification may be ruled out in other applications. If classification has to be automatic, it must be based on an ontology. Categories must not only be well defined, their discriminating features must be made explicit. This will enable the user to exploit the semantics of the domain to navigate efficiently to the desired products.

Principle 6: Pragmatism: KO for the Web is determined by the needs of the Web-based application.

Finally, all other aspects of the KO scheme must be determined with pragmatism as the key. Should a category be present? Should it have an alternative name? What kind of links should be there among knowledge elements? All such questions must ultimately be decided based on the needs of the particular Web-based application.

Knowledge Organization for the Web: Superlinks with Semantic Roles

Although the above principles highlight the importance of ad-hoc and user-specific categories, it does not mean that we can simply do away with knowledge organization and allow social tagging and folksonomies (Lambe, 2007) to take their place. While social tagging and democratic review and rating processes are useful and must be included in the overall scheme for KO, social anarchy cannot be the model for organizing knowledge. There is no guarantee that such a loose scheme can actually deliver acceptable results. At the same time, the variety of KO schemes that have been proposed for traditional as well as Web-based applications (Garshol, 2004), such as decimal and faceted classification, Dublin Core meta-data, topic maps, concepts maps, RDF, and so on, also do not fully meet the requirements of a Web-based application.

Elsewhere, we have demonstrated that conventional hyper-linking on the Web is inadequate for semantics and KO purposes and proposed a new kind of semantic, multi-way linking known as Superlinks (Mahesh and Karanth, 2012; Karanth and Mahesh, 2012). We show here that Superlinks are ideal as a KO scheme for Web-based applications. A Superlink is n-ary in nature in that we can link one source to multiple target locations while also capturing the context in which link is relevant by specifying the semantic role for each target. The source and destinations of the link can be at any level of granularity ranging from a single word or phrase to a paragraph, section or the entire document itself.

Consider a Web-based e-commerce site for selling cellular phones along with their accessories and software applications. The site has to make it easy for users to match specifications of cell phones with those of accessories to determine compatibility between
an accessory and a particular phone model. There is no ready-made set of subject headings or a standard set of facets for a domain like this. Nor is the content made up of well-edited, self-contained or properly interlinked documents. What we have is a mix of product-related documents (e.g., specification sheets), data in the form of tables and charts, and much unstructured content in the form of customer reviews, questions and comments. Nevertheless, the richness of the domain as well as the context of a user’s interaction with the site must be captured by the KO scheme applied to organize all of this content in a manner most suitable for the user.

We propose that a KO scheme for this application must have these features:

- A domain ontology that interlinks all of its key entities and concepts (e.g., cell phone, operating system, application, capacitive touch screen, battery life, etc.);
- A linking mechanism that captures the semantics of the links;
- A multi-way and precise linking mechanism to interconnect several related pieces of text at a fine level of granularity; and
- Means for enabling users to comment, review and rate any part of the KO.

We build an ontology in OWL language covering all cell phones present in the Web-based store, their attributes like type, operating system and battery life, and features such as camera, MP3 player, Bluetooth, Wi-Fi, GPS and USB as well as different accessories for the phones. Detailed specifications of various phone models and accessories are available as semi-structured Web pages and PDF documents.

How do we create a linking mechanism that permits complex, semantic linking between concepts in the ontology and precise locations in associated texts (i.e., Web pages) as well as among different places across Web pages? Clearly, the existing hyperlink structure is inadequate since it neither allows multi-way links nor any semantic attributes of the links. We propose a type of link called Superlink that is built using the XML family of technologies including XSD, XLink and XPath (Harold and Means, 2001).

The key to establishing the context in describing the knowledge resources on the Web site is to link relevant concepts in the ontology to their sources in textual documents. A “Superlink” captures the semantic relations between concepts and pieces of text by specifying the semantic role played by the link. An application-specific typology of semantic roles can be employed, for example, definition, illustration, example, compatibleWith, etc. Superlinks with such semantic roles permit ad-hoc, impure and user-specific ways of linking source and targets across the Web site.

Figure 1 gives a pictorial representation of how Superlinks are created between specifications of phones and accessories. It shows a Web page about a cell phone that is linked together to two locations in two different pages where compatibility of the phone with related accessories is mentioned. A particular cell phone’s specification can be linked to multiple accessory specifications based on the semantic role CompatibleWith. The example shows how “HTC Wildfire S” specification can be linked to its multiple accessories like “HTC Portable Bluetooth Conference” and “Stereo Headset” using a single Superlink with CompatibleWith as the semantic role between the source and the two targets.

Figure 2 shows an XML encoding of the Superlink that includes two locator tags along with attributes such as xlink:show and xlink:actuate that specify what actions the Web browser should take when the Superlink is clicked. In addition, xlink:role specifies the semantic role of the link between the source page and that particular target which is itself linked by xlink:href, the target URI.
Fig. 1: Three Web pages linked together with a Superlink

Semantic Role - "Compatible with"

HTC Wildfire S
- Live it. Love it. Share it.
- Size: 101.3mm x 59.4mm x 12.4mm 3.99" x 2.34" x 0.49"
- Weight: 105 grams (3.7 ounces)
- Display: 3.2-inch touch screen

Description
- Compatible two people are a cold world, but the Bluetooth Conference Speaker connects by your phone to other Bluetooth-enabled devices making your mobile office a music phone range.
- It's easy to connect to the mobile user and call 2 devices. Then you can listen to your favorite music.

Technical Details
- Talk time: 10 hours
- Standby time: 300 hours
- Recharge time: 2 hours
- Country of manufacture: Korea
- Country of Design: Korea

Also Compatible With:
- HTC Wildfire S
- HTC Desire HD
- HTC Desire Z
- HTC Desire
- HTC Wildfire
- HTC Magic
- HTC Legend
- HTC Aria
- HTC Hero
- HTC Desire

Fig. 2: Superlink with multiple locators for CompatibleWith relation

<Superlink xlink:type="extended" xlink:show="new" xlink:actuate="onLoad">
  <Locator xlink:type="locator" xlink:show="new" xlink:actuate="onRequest">
    <h1>HTC Wildfire S</h1>
    <p>Live it. Love it. Share it.</p>
    <ul>
      <li>Size: 101.3mm x 59.4mm x 12.4mm 3.99" x 2.34" x 0.49"</li>
      <li>Weight: 105 grams (3.7 ounces)</li>
      <li>Display: 3.2-inch touch screen</li>
    </ul>
  </Locator>
</Superlink>

Semantic Role

Source of Superlink

Targets
Conclusion

In this paper, we have summarized the knowledge organization requirements of Web-based applications in a set of six principles. Further, we have proposed ontology-based semantic Superlinking as the KO solution for applications such as Web-based e-commerce. Further work is needed to solve the challenges of engineering large-scale solutions based on this framework. Attempts must also be made to integrate such solutions across multiple Web-based services so that a common ontology as well as cross-links can be developed across compatible services.

References
Gercina Ângela Borém de Oliveira Lima - Federal University of Minas Gerais, Belo Horizonte, Brazil


Abstract
This research is focused on the continuation of the Hypertext Map prototype implementation – MHTX, proposed by Lima, (2004), with the general objective of transforming the MHTX into a semantic content management product facilitating navigation in context supported by customizable software that is easy to use, through high end desktop/web interfaces that sustain the operation of its functions. Besides, these studies aim, in the long run, to achieve the simplification of the information organization, access and recovery processes in digital libraries, making archive management by authors, content managers and information professionals possible.

Introduction
Directing users effectively to the information that circulates on the Web is a problem that has concerned many researchers. According to Lancaster (1994), "in thirty years we went from the inexistence of databases in electronic format to the reality of thousands of them". Today, the information demanded by the user could be in any format, and be located in any part of the world. Lucas (1995, p.69) reminds us that the new information technologies may be useless without proper means to filter, organize and summarize their products. Thus, we can assume that if a piece of information is made available in an organized and logical manner, it will be easier to locate and recover what is being searched for. Ideal texts would be semantically organized, with a conceptual structure reflecting the semantic relations between their subjects.

Hypertext, with its special information storage and recovery characteristics, may offer solutions to meet contemporary demands for information. However, given the enormous quantity and plurality of documents available on the Web (with its dispersion of products and producers), the regular and precise construction of semantically structured hypertexts requires the development of adequate methodologies, applicable to different types of texts. Since the advent of hypertext documents in the 1980s, the literature has pointed that navigation problems are caused by the absence of clear and consistent criteria in the conceptual organization of hypertext. The main problems occur in the phases of choice and structuring of relevant information and the creation of pertinent links.

Most of the time hypertext is planned and structured by chance, without considering the semantic structuring of the document or the needs of potential users. According to Schiper¹, quoted by Lucas (1995, p.70), hypertext should be structured according to the logic of knowledge and the principles of reasoning within the dynamics of text reception, in which a text may receive various readings and interpretations.

The proliferation of products that use techniques of document hypertexting – such as sites, digital libraries etc, has not been accompanied by the creation of rigorous methodologies. The creators of hyper-documents are generally untrained people who learn to manipulate hypertext construction software, without knowledge to semantically structure textual information.

A disorganized navigation trail, resulting from a conceptually deficient hypertextual document design tends to present obstacles in searching and browsing that inhibits recovery through search engines.

Objectives and Methodology

This paper succinctly presents the methodological approaches and findings related to the implementation of the Hypertext Map Prototype (MHTX) by the research group coordinated by Lima (2004) at UFMG (Brazil). Generally speaking, it aims at creating a hypertext management product that uses semantic content with context navigation from any area of knowledge. It eventually envisions the construction of friendly and customizable software through high-end desktop/web interfaces.

Specifically, it intends to:

- help the organization and representation of human knowledge in hypertexts, based on semantic referential structures through hypertext links;
- study the contribution of specific automatic indexing techniques with a faceted taxonomy as target, based on electronic sampling of academic document content;
- propose the transformation of faceted taxonomy into an ontology using the XML/OWL standard to facilitate navigation in context within electronic documents. Eventually, it would generate a semantic tool to support information representation in the MHTX Database;
- study the applicability of the Topic Map Standard in the compatibility of hypertext maps arising from the generated taxonomy; and
- propose a faceted semantic interface for representation and navigation in hypertext maps.

Our approach to the modeling of hypertext documents considers the following assumptions and respective methodological references:

a) A hypertext system (Conklin, 1987) does not reach a desired level of effectiveness in access and recovery if the hyper-documents that comprise it are not conceptually structured.

b) Automatic indexing may assist the process that precedes conceptual categorization.

c) The Theory of Facet Analysis (TFA), advocated by Ranganathan (1967), may present great potential to logically structure hypertexts.

d) The Concept Theory, proposed by Dahlberg (1978), may be used to semantically structure hypertexts.

e) The principles for the construction of Conceptual Maps, (Novak, 2006) which emerged in the education field may be used in the operationalization of the proposed model.

f) The Topic Map Standard may be instrumental to make conceptual maps compatible among themselves.

This research is linked with Master’s and Doctoral work under the supervision of the present author and is related to the question of semantic content management raised in the attempt to convert the MHTX into customizable and friendly software. So far, the research to address the issues includes:

1 analysis and comparison of software hypertext interfaces for the construction of hierarchical diagrams, conceptual maps and hyperbolic maps of their characteristics and analysis of resources that allow for the representation of knowledge through a system of concepts;

2 studying the compatibility of conceptual maps arising from theses and dissertations located in the UFMG Digital Library in Brazil and studying the applicability of the Topic Map standard, especially in relation to the merging process of these maps, using compatibility language theories developed by Ingetraut Dahlberg (1983), Dagobert Soergel (1982) and H. H. Neville (1970);
studying automatic indexing using extraction of relevant terms for the representation of informational content, based on criteria similar to those utilized by human beings;

studying the existing relations between hypertext systems, through the link determining process and the thematic representation process based on the concept theory (Dahlberg, 1978);

studying the relation between the theories of categorization (classical and contemporary cognitive) and Ranganathan’s principal categories, treating categorization as a cognitive process, as well as understanding the different categorization instances in Library and Information Science;

analyzing of the ontology construction process, investigating strategies for the development of methodological formalization with the objective of standardizing this process from the Information Science point of view;

studying the contributions of usability principles in the UFMG Digital Library interface;

the proposal of a navigational faceted taxonomy departing from a category matrix developed for academic digital texts, aimed at a systematized and semantic representation to help search and recovery;

propose an architecture for semantic digital libraries based on theories developed over the years for representation and organization of knowledge in traditional, as well as recent possible advances made by technology developed in the context of the semantic web and the social networks; and finally; and

studying the organizational potential of metadata in digital libraries with the objective of constructing software capable of carrying out the process of managing and recovering scientific virtual documents.

Findings

The optimization process of the MHTX Prototype yielded the following projects and findings (in chronological order):

- Silva (2007) presented a comparative analysis of hypertextual interfaces with respect to the capacity of software to represent and retrieve information. It compared three types of hypertextual interfaces — hierarchical diagrams, conceptual maps, and hyperbolic maps — revealing the advantages and disadvantages of these interfaces in each application. The parameters for the analysis of conceptual representation capacity were based on theories and methodologies from the Information Science area, namely faceted classification theory, concept theory, the ISO 2788 standard for the elaboration of monolingual thesauri, and the ISO 704 standard for terminological practice. With respect to capacity for information retrieval, the available resources to facilitate browsing in concept systems were mainly assessed. Findings show that each one of the interfaces had special advantages:
  1) the hierarchical diagram with explicit representation of relations;
  2) the hyperbolic map with its fisheye feature, which allows the representation of a knowledge domain in a more detailed form while keeping control of the whole and the parts at the same time.

Finally, it should be emphasized that these interfaces may coexist rather than in isolation.

- Silva (2008) studied the use of topic maps, with the main objective of automating the process of making these structures compatible, rather than the traditional
representation in conceptual maps. Silva’s idea was to make possible the fusion among topic maps in order to promote a semantic inter-relation between conceptual maps and, consequently, between the hypertextual information resources. Nevertheless, this process presented great complexity and required a high level of manual interference. The fusion results were analyzed in the light of the attempts to make the theories of indexing languages developed in the area of Library and Information Science from the 1960s onwards compatible. The principal objectives achieved were:

1) Detailed conceptualization of the topic map fusion process, considering the compatibility levels possible and the applicability of this technology in the integration of faceted structures;

2) Detailed sequencing of steps that may be used in the implementation of topic maps, based on faceted structures.

One of the problems in technologically implementing the MHTX is the prohibitive cost of manually indexing large collections, such as all the dissertations and theses to be included in the database. Borges (2009) studied automatic indexing. Based on previous studies on this subject and using relevant literature published between 1950 and 2008, various relevant criteria were identified to compile a list whose combination would supply the most satisfactory results for the construction of parsers to handle textual documents and for digital libraries, more specifically in the implementation of the MHTX Prototype. Findings include the selection of nine criteria considered essential for the development of automatic indexing software, namely:

1) Word phrase formation,
2) Absolute frequency of the occurrence of the word in the text,
3) Word identification (Comparison with use of dictionary),
4) Identification of Word stemming,
5) Stopwords,
6) Numerical weighting,
7) Term weighting,
8) Semantic vocabulary / conceptual heading, and
9) Vocabulary/Thesaurus.

In the original version of the MHTX, Ranganathan’s Theory of Facet Analysis (TFA) was used for conceptual modeling of hypertext systems. However, there was a need to study new semantic relationships that would allow for the connection among the conceptual links of the hypertextual academic document in order to yield a more comfortable navigational architecture. It would provide users with browsing possibilities that were not simply established through hierarchical and associative relations. Nonato (2009) studied the application of the concept theory proposed by Dahlberg (1978) in a link sampling of the MHTX prototype, departing from the application of facet analysis. Findings showed an increase in the amount of relationships between concepts, which made the link labeling process clearer. Thus, the choice of a more suitable relationship could be established, taking into consideration the context the link would be remitted to. Finally, it was also observed that the information organization process – through concept theory – offers theoretical possibilities that make explicit the systematization between concepts, such as its organization in the memory of hypertext users.

In order to broaden the categorization possibilities and help the elaboration of controlled vocabulary to be used in the MHTX System, Silva (2010), through
exploratory research, resorted to elements essential to this subject and analyzed the categorization process in Library and Information Science (LIS), performed under the scope of Ranganathan’s principles. It aimed at bringing the theoretical debate within the empirical process. To understand the relation between categorization theories (classical and contemporary cognitive) and Ranganathan’s categorical principles, it departed from the assumption that Ranganathan’s categorical model predominates in LIS and that this model was based on the writings of Aristotle. Furthermore, in the different instances of categorization exercised by professionals, the categorical processes occur in a distinct manner. The study revealed the difficulties professionals had in applying the PMEST structure, which may be related to the distance between these theories and the professionals’ daily work, since they don’t use the PMEST structure explicitly in the execution of LIS tasks – except in IRI drafting. Thus, there are indications of the use of the PMEST structure, as a categorization process guide in an unconscious manner or, if conscious, performed in a really subtle and loose way. It could be inferred that the categorization process in the different instances of LIS – the classificationists and the classifiers – applied to the universe studied, reflects a weaken appropriation of the PMEST scheme.

Demarques (2011) did an exploratory study, complementing the procedure for making structures compatible with that proposed by Silva (2008; see above) through using a domain ontology in the MHTX database, seeking reflections on the reading and writing practices of documents published in digital libraries. In these environments, the ambiguity of the vocabulary was identified as one of the crucial aspects for the semantic compatibility of heterogeneous bases. The proposal, aiming at an MHTX enhancement, was to analyze the representation of different publications in a single map as a system of integrated concepts, based on the methodological-theoretical foundation of vocabulary compatibility. An informational digital object modeling technique was proposed, considering the same principles of faceted classification. This technique allowed the use of concept functionalities to explain equivalence relations in a single map, according to assumptions that consider ‘concept’ a descriptive, administrative and structural metadata. To this end, both indexing and retrieval must be based on a consensual terminology for the establishment of formal categories directly linked to the use of language. The combination of these categories represented in facets of a domain ontology showed semantic connections in which the characteristics expressed in the hypertextual objects are present in the definition and in the relations between these hyperlinked objects. Results confirmed the feasibility of inserting a domain ontology in the MHTX, utilizing the proposed modeling with automated indexing mechanisms for semantic interoperability and compatibility, which may help the interactive retrieval of information.

Pereira (2011), continuing Silva’s 2007 study, did a case study on the usability of digital library interfaces, verifying what type of interface would be mostly suited to the MHTX Prototype. The scope covered the Digital Library maintained by the IBICT, using this library’s interface as a sample. Two assessment methods have being used so far: the heuristic evaluation for usability proposed by Nielsen (2005) and the empirical assessment method. Findings include a list of the usability problems that compromise user interaction with the system; specific solutions to these problems were proposed.

Maculan’s research (2011) on a faceted browsing taxonomy for academic works aimed at structuring dissertation and thesis type documents into nine Basic Thematic Classes (BTC):
1) Theme;
2) Object;
3) Environment;
4) Type of research;
5) Data collection;
6) Methods;
7) Theoretical foundation;
8) Historical/contextual foundation; and
9) Results.

Besides, he developed an algorithm for a sequence of tasks to systematize conceptual analysis for the extraction of concepts in theses and dissertations. This algorithm, called “Thematic categorical matrix for academic works” (Maculan, 2011), uses data representation in three columns:

**Column 1:** the BTC set;

**Column 2:** “queries”, as suggested by NBR norm 12676:1992 and by the analysis principles of the PRECIS indexing system (Fujita; Rubi, 2006);

**Column 3:** identification of documents’ textual structure parts where the answers to the second column queries can be found.

Review of the literature showed that faceted browsing approach (with categories) is being used on the Web since the end of the 1990s in exploring, searching and retrieving information. This study showed the feasibility of a browsing taxonomy in specialized services, such as digital libraries of theses and dissertations. It was possible to observe that the principles of facet analysis can be used in the preparation of faceted classifications or taxonomies, both for the indexing of content and as a mechanism for navigation, facilitating information search and retrieval in digital libraries. Faceted taxonomy has been consistently efficiently used within the MHTX project, from data entry in the system to yielding faceted interfaces for searching and localizing information.

- A study of architecture models for semantic digital libraries is in progress.
- Usage of metadata in digital libraries, which complements the study by Demarques (2011), mentioned above is also in progress.

**Final considerations**

The main goal of the MHTX research group is, in the long run, to achieve the simplification of information organization, access and recovery processes in academic digital libraries, making easier archive management by the authors, content managers and information professionals. These studies also aim at contributing to the development of research on semantic organization of texts, based on cognitive analysis followed by the corresponding technological application. Finally, it intends to encourage interdisciplinary research fronts that integrate Information Science with areas such as Computer Science, Cognitive Science, Linguistics and Education.

**Bibliographic References**


Silva, Alessandra Rodrigues da. 2010. Estudo dos princípios de categorização na Biblioteconomia e Ciência da Informação: Ranganathan - entre a teoria clássica e a abordagem cognitiva. (Mestrado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte.


Evelyn Orrico - Federal University of the State of Rio de Janeiro, Brazil
Vera Dodebei - Federal University of the State of Rio de Janeiro, Brazil
Miriam Gontijo - Federal University of the State of Rio de Janeiro, Brazil

The Precision of Metaphor for Information Retrieval

Abstract
We present a theoretical model for an information retrieval filter based on metaphors. The justification for this proposal is the current information boom, especially due to new technologies. The theoretical basis for our proposal is an association of the Cooperation Principle proposed by Grice (1975), the Indolence Principle by Berrendonner (1989), Lakoff and Johnson's (1980) concept of metaphorical representation, and Ranganathan's (1967) faceted system. A case study was done with a group of researchers in the field of Transport Engineering, as well as an analysis of manifestos written by a group of contemporary artists with the purpose of broadening the scope of the proposed model. We conclude that the model is feasible, although it would benefit from further studies involving different fields of knowledge.

Background
Bouramoul’s (2011) paper on the semantic dimension in information retrieval, taking Query Reformulation document indexing as a starting point, led to the resumption of an investigation carried out a decade ago (Orrico, 1999) in which a theoretical model was conceived for information retrieval based on metaphors. At that time, the research was motivated by the perception that contemporaneity brought many technological conveniences for the dissemination of information. However, such conveniences have also caused a build-up of information material that demands great effort in order to appropriately handle the large number of resources on the web. This perception is becoming increasingly strong these days. Besides the quantity, another problem arises due to the ease resulting from the use of new technologies: an ever-increasing universe of users may have access to wider and more varied information, which may potentially be inappropriate for them.

Thus, this proposal seeks to resolve, in principle, two problematic situations: a) the difficulty in highlighting, among a vast set of possibilities, the information that really matters; b) limiting access to inappropriate information for some segments of the population.

Considering the assumption, as proposed by Lakoff and Johnson (1980), that the human mind represents the world through metaphors, the fields of knowledge would also be represented metaphorically by those devoted to them.

In fact, using metaphors to represent knowledge is not new. Orrico and Oliveira (2005) showed that arborescent representation gained popularity in the 16th century. Because of that, Ramon Llull’s Arbor Scientiae, written around 1300, was reedited many times, and several other representations were elaborated in different areas: grammar trees, love trees, logic trees (the Porphyrian Tree), Jesuit’s trees etc. During the Age of Enlightenment, allied to this arborescent representation, the task of organizing and classifying knowledge executed by encyclopaedia authors reveals the use of another metaphor: the world map. This is how d’Alembert describes, in the Encyclopédie, the difference between an encyclopaedia and a dictionary:

[like] a type of world map, which must show the main countries, their position and their mutual dependence, the roads that lead directly from one to another. This road is often interrupted by thousands of obstacles, which are known, in each country, only by the travelling inhabitants, and cannot be represented except in highly detailed individual maps. Such individual maps will be different entries in the Encyclopédie, and the tree, or systematic map, will be its world map. (Darnton, 1986, p. 251).
In the context of that era’s project, which relied on various strategies, knowledge thus represented was transferred to intellectuals and followed a path that resulted in the secularization of education and the emergence of academic disciplines. As a consequence of the development of specialization, the ideal of the “universal person”, whose erudition was not limited to a single field but covered all types of knowledge, was attenuated. (Burke, 2003, 82a)

Therefore, the polarization between specialization and generalization arises from the fragmentation of knowledge, a development which, at least on some occasions, was regretted. As Burke (2003, 81a) says: “General knowledge became necessary due to the connection among things and interdependence of notions, such that one piece of knowledge sheds a light on the others”.

To answer the proposed questions, we selected a group of researchers from the Transport Engineering field, whose specific object —means of transportation— is studied by various areas. Our intention was to identify how each researcher represents his/her own field of knowledge and investigation, considering that the researchers in the group come from different academic areas. Moreover, knowing that this group included researchers with varied levels of experience, we assumed that the representation of the field would be constructed during the development of activities by the group.

By means of questionnaires sent by e-mail to each group participant, we could observe the metaphoric phrases used to designate their research field. Based on the analysis of these answers and on the identification of noun phrases used to designate the field, we could see that the metaphors used by the researchers were similar. This allowed us to infer that they shared a common representational universe and to conceive the proposal for the model.

Aiming to validate the theoretical framework of the proposed model, another investigation was carried out (Coelho, 2006) in a different field, this time Contemporary Art, and particularly Conceptual Art. This study was used only to reinforce our certainty about the theoretical framework.

Objective
The objective of this text is to present a theoretical proposal for an information retrieval filter based on metaphors used in natural language.

First we present the current scenario of excessive information in circulation; then we introduce the theoretical framework that supports the arguments behind our proposal, the case study, and finally the proposal itself.

Context
As predicted by Solla Price (1965), academic production grows exponentially, doubling its production every ten years. What can be said, then, about this new space of information dissemination that is the Internet? Context shows that Orrico’s (2002) proposal, first presented for discussion in this ISKO forum during its meeting in Granada, is still valid. Raieli (2012) even points out the broader problem of aggregating documents of different natures for retrieval.

Due to Internet’s internal organization, this means of virtual communication allows a virtually unrestricted amount of data supply and access, which can cause, as pointed out by Pierre Lévy (1999), what Roy Ascott termed “second deluge”, which is characterized by the “chaotic and exponentially explosive nature of telecommunications growth”.

This scenario is by no means new and information professionals seek increasingly creative alternatives, making use of natural language to enable access, search, and retrieval of information circulating on the web. Current studies of ontologies are evidence of this.
According to the initial proposal by Guarino (1998), formal ontology starts off from a specific description of the objects to be retrieved, establishing a relation of dependency between them and, according to Park (2008), the ontologies have been cited as a critical part of the projects of information systems, to a great degree because they help sustain a communication scheme in the fields of interest of a certain social group.

It is assumed that the proposal presented here can contribute to the studies of ontologies because it is part of a scenario which takes context into account in order to determine meaning. For this assertion, it bases itself on Durkheim’s (apud Burke, 2008, 78) words, according to whom —The categories of human thought are never permanently fixed; they constantly build, unbuild and rebuild themselves; they change with place and time”.

To achieve our broader objective, that is, helping a given group retrieve the information that interests it the most, this article is based on a theoretical semantic framework, specifically on metaphors used by certain social groups to represent both the area of knowledge in which they work and the concepts connected to it.

In order to assume that metaphors are adequate for information retrieval, this article works on the assumption that meaning is built up within an interactive-communicative process, whose organisation is governed by principles of communication. The principles underpinning our analysis are Cooperation, proposed by Grice (1975), and Indolence, by Berrendonner (1989), as shown in Orrico (1999). These principles allow us to relate Lakoff and Johnson's (1980) concept of metaphor to Ranganathan's (1967) revolutionary faceted classification system for the classification of knowledge.

**Theoretical-methodological framework**

For Paul Grice (1975), human communication is organised according to the Cooperative Principle, whose assumption is that individuals who communicate build utterances according to four maxims: Quality (Truth), Quantity (Information), Relation (Relevance) and Manner (Clarity). According to the first maxim, it is assumed that everything the interlocutor says is true; the second maintains that he/she only says what is necessary; the third, he/she only says what is relevant for communication and, finally, he/she does it in the best possible manner.

Going a step further from the previous principle, Berrendonner’s (1989) Indolence Principle establishes an interface between communication and cognition insofar as it says that in utterances there is little informative content. It is possible to formulate little informative material because the speaker takes into account that the listener already has enough information which enables him/her to infer further specifications from an utterance with underspecified content.

The interaction of metaphor with these two principles occurs because metaphor is a figure of speech that transfers a term to a scope of meaning outside its own scope in order to establish a representation of the world by means of analogies. These analogies, built within communication processes, follow the four principles presented in the previous paragraphs.

Metaphor is here understood from Lakoff and Johnson’s (1980) point of view, where they maintain that human beings organise knowledge making use of structures called idealized cognitive models [ICMs] and that categorial structures are derived from that organization. The proposal of these models admits that mental organisation occurs through the cultural construction of schemes of knowledge about the world, through the construction of what authors call ontological metaphor, which is the core metaphor from which others derive, and which help this representation. To the extent that metaphors are used to represent socially constructed categories, let us then discuss what meaning representations are.

Meaning representations refer us to the father of library sciences, S. R. Ranganathan. His idea of classification, which is both innovative and has important repercussions for the
theory of classification as a whole, reflected the modern concept of knowledge which persisted at the end of the nineteenth century and the beginning of the twentieth century, and which presupposed a precise delimitation of the area of knowledge where such concepts existed. Although nowadays we deal with imprecise limits to understanding cultural phenomena, Ranganathan’s faceted proposal builds an adequate framework within which metaphors are found.

**Case Study**

This theoretical dialogue was an important support for a case study, carried out with a team from the transport area in Brazil, consisting of twenty-eight workers with different levels of experience and from various regions of the country. Data collection was carried out by means of a questionnaire, electronically sent and answered, and its main aim for the writing of this paper was to generate representations of the transport area by analysing the following question: “What exactly do you understand by transport?”

As a methodological analysis, sentences were analysed in which the interviewee explicitly used—or left no margin of doubt about—the verbs *be* or *represent* in utterances of the type “Transport is…” or “Transport represents…”

The analysis of the replies revealed a recurrent use of terms which can place transport within three semantic fields: a) network / set / system; b) movement / dislocation; and c) manner / way / procedure.

The proposal of the theoretical model identified the relationship between this reply pattern and Ranganathan’s faceted categories, assigning to the personality category the ontological metaphor that represents the area being investigated. This metaphor was selected as ontological because it was used by most research participants and was present in the discursive context indicating the nature of the field.

Other representations appeared in the universe of answers in discursive contexts referring sometimes to the role played by the Transport field and sometimes to the conditions or manners in which it operates, with no representations related to the space and time in which Transportation occurs.

This fact indicates that, maybe because these dimensions are inherent to the Transport theme, there is nothing to be said as specific representations, and thus it is not pertinent to draw attention to them. Taking this into account, these categories were considered secondary in the creation of the filter, since, in a model intended to filter information, the focus must be on what determines specificity rather than similitude.

Based on these observations, we established a correlation among the representations of the transport field made by the researchers and the faceted categories proposed by Ranganathan. This correlation lead to Table 1, where the first column presents Ranganathan’s categories, the second presents the correlation with the proposed model, and the third presents the nuclei of the noun phrases used by the researchers to represent the field under investigation.

**Table 1: Categories of the Transport Model**

<table>
<thead>
<tr>
<th>Ranganathan’s Categories</th>
<th>Model Categories</th>
<th>Metaphors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fundamental Personality</td>
<td>Essence</td>
<td>Systemic network</td>
</tr>
<tr>
<td>Matter</td>
<td>Function</td>
<td>Movement</td>
</tr>
<tr>
<td>Energy</td>
<td>Conditions or Manner</td>
<td>Procedure</td>
</tr>
<tr>
<td>Secondary Space</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
We believe that, based on the metaphorical representations, it is possible to identify those that represent the essence—or personality—of the field and the other categories, as proposed by Ranganathan, in order to create a user-oriented information retrieval filter.

This conceptual framework that supports the filter must focus on aspects of the user’s cognitive behaviour from a qualitative perspective, since it has to consider, apart from the field of knowledge in which the search will take place, the search strategies to be used and the goals of the user.

Using a filter would allow a researcher to retrieve the information s/he considers most relevant according to his/her universe of knowledge, optimising the searches and achieving a good satisfaction level.

As was mentioned, an investigation was conducted in another area, now not among scientific fields but in art. The study had the goal of extending our analysis to a different set of users in an attempt to investigate other discursive practices that could indicate inconsistencies in our model. The purpose was not to exhaust the discussion, but only to broaden the scope of the theoretical reflection.

In the art field, we turned to Contemporary Art, specifically a cultural movement called Conceptual Art. It is common knowledge that this type of art could only be appreciated by an experienced spectator, in the sense of someone who is already familiar with the symbolic universe that governs this type of art. In our study, we tried to identify the discursive construction of the metaphorical representations that the conceptual artists elaborated in their artistic essays/manifestos, aiming to reveal the representations used to establish the identity of that group.

In this group devoted to art rather than science, we again identified the regular use of metaphorical representations in the production of the group of artists, and that resulted in Table 2:

<table>
<thead>
<tr>
<th>Ranganathan’s Categories</th>
<th>Model Categories</th>
<th>Metaphors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fundamental Personality</td>
<td>Essence</td>
<td>Action</td>
</tr>
<tr>
<td>Matter</td>
<td>Function</td>
<td>Artistic expression</td>
</tr>
<tr>
<td>Energy</td>
<td>Conditions or Manner</td>
<td>War / Fight</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Secondary</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Space</td>
<td>Space</td>
<td>Brazil</td>
</tr>
<tr>
<td>Time</td>
<td>Time</td>
<td>Military regime</td>
</tr>
</tbody>
</table>

In the specific case of this type of artistic manifestation, although Space and Time categories continue to be secondary, they need to be explicit because, in the art field, especially during the military regime, the fact of being in Brazil was something that needed to be signalled.

Based on these two studies, we believe that it is possible to broaden the reach of our model.

**Conclusion**

Any and all classification criteria, with their intrinsic relevance rationale, ultimately aim to meet the expectation of user satisfaction. In this case the proposal aims to use metaphors which represent categories that comply with the necessary representation conditions for this area of knowledge. This classificatory organisation aims to increase the accuracy in meeting the demands in the context of new information technologies. Metaphors in this case would help attain higher accuracy, however paradoxical this may seem at first. In fact, we regard metaphor less as a literary style feature which points towards inaccuracy and more as a degree of representation which points towards specifying concepts used in a given area of knowledge.
References


Charles-Antoine Julien - McGill University, Canada
Pierre Tirilly - University of Wisconsin, Milwaukee, USA
John. E. Leide - McGill University, Canada
Catherine Guastavino - McGill University, Canada

Using the LCSH Hierarchy to Browse a Collection

Abstract
The Library of Congress Subject Headings (LCSH) is a subject structure used to index large collections throughout the world. Browsing a collection through LCSH is difficult using current on-line tools in part because they are inadequately integrated with information collections. Users of these LCSH browsing tools are expected to find a promising LCSH string before using it to search for the information itself; many users do not have the patience for such a two-step process. This article proposes a method to fully integrate a specific collection in its subset of the LCSH hierarchy in order to facilitate LCSH browsing as well as information retrieval. Techniques are described to match LCSH strings assigned to the collection with an established string from the authority records, and build their specific LCSH hierarchy. The resulting subset of LCSH structure is described in terms of its size and broader/narrower term statistics, and implications for browsing and information retrieval are discussed. The results of this research have implications for institutions wishing to further capitalize on existing LCSH organization investments for the purpose of subject browsing and information retrieval.

Introduction
The Library of Congress Subject Headings (LCSH) is a controlled subject vocabulary which covers all knowledge domains and is used extensively throughout the world. However, the use of LCSH in digital search environments has been very limited in part because of the inadequacy of current online tools for browsing collections using their LCSH structures. Representing a collections’ subset of the LCSH structure as a hierarchy is a promising approach for improving subject browsing (Zhu & Chen, 2005). This study adds to the work of Yi and Chan (2008) by considering how a real-world bibliographic collection is distributed within the LCSH structure, and exploring this structure’s potential for subject browsing.

Current Library of Congress Subject Headings (LCSH) browsing systems do not provide explicit features to navigate the LCSH structure, forcing users to visit each LCSH individually. Traditional alphabetical displays of LCSH has long been recognized as inadequate (Drabenstott & Weller, 1996) since it hides the structure of interrelated concepts (Larson, 1992; Richmond, 1959). Inadequate LCSH browsing features may partly explain why its "hierarchical nature is largely ignored" (Frank & Paynter, 2004) by the searcher.

The few existing LCSH browsing interfaces (Papadakis, Stefanidakis & Tzali, 2009; Yi & Chan, 2008) are designed as interfaces to the empty LCSH structure which does not contain a bibliographic collection. This article differs from those prior works since it restricts the LCSH structure to contain a specific collection so that searchers are able to browse the information based on its information organization structure. Specifically, we:
- offer an automated method to reduce the LCSH structure to reflect the contents of a real-world collection;
- verify the presence of known LCSH structure issues (Shubert, 1992; Yi & Chan, 2009);
- provide some mitigating solutions to these structural issues so as to improve collection browsing based on its LCSH hierarchy;
- confirm the presence of a power law distribution of information within the LCSH structure (Bates, 2003).

1 http://authorities.loc.gov and http://www.oclc.org/connexion
The LCSH Structure

A bibliographic collection organized using LCSH is comprised of two data sources: 1) bibliographic records that are indexed by topic with at least one topical LCSH string, and 2) the LCSH authority records that contain the established topical LCSH strings and define the topical LCSH hierarchy. A topical LCSH string is comprised of a topical main heading, and possible subdivisions meant to further specify the main heading. The following definitions are used in this article:

- **LCSH string**: any topical LCSH
- **Established (LCSH) string, Authority record**: a LCSH string listed in the Library of Congress (LC) authority records. LCSH strings not listed in the LC authority records (i.e., non-established strings) are subdivided strings created by cataloguers without being added to the authority records.
- **Main heading**: the left-most portion of a LCSH string indicating the main topic of the string, excluding any subdivisions. Note that every LCSH string has a main heading, and a string without subdivision(s) is equal to its main heading.
- **(LC) Authority records**: the list of established LCSH strings. This includes all valid main headings and some subdivided main headings. Authority records define the relationships between main headings (i.e., broader/narrower or related terms).
- **LCSH (syndetic) structure**: the sum of relations between main headings as defined in the authority records.

Main headings can be related with other main headings in one of two ways: in a hierarchical relationship between a broader term and the reciprocal narrower term(s) (BT vs. NT) or in an associative relationship between two related terms (RTs) (Chan & Hodges, 2007). Broader terms are vertical relationships that define a hierarchy of broad (e.g., Science, Medicine) to narrow terms (e.g., Mechanics, Surgery). Related terms represent horizontal relations; these are not considered by the current study.

The broad to narrow relations between main headings define a LCSH topical hierarchy which has known structural issue. Firstly, there is a large number of orphan branches that have no broader term (see review by Shubert (1992); (Yi & Chan, 2009)). This causes a usability issue because users browsing the first level of the hierarchy are asked to choose from a long list of unconnected branches (Yi & Chan, 2008). Secondly, main headings often have multiple parents; this entails that the LCSH structure is a lattice (Besag, 1974) which is difficult to browse. The extent of these structural issues is described in the following section.

**Method**

This section first describes the bibliographic and LCSH authority data sources used in this study. This is followed by the extraction of individual LCSH strings assigned to the bibliographic records. These are matched with their corresponding established LCSH string to retrieve the broader relations that define the LCSH hierarchy. LCSH structural issues concerning orphan LCSH strings are mitigated, and the result is the LCSH hierarchy containing the collection to be browsed. Finally, this collection specific LCSH structure is described in terms of its size, and numbers of broader/narrower terms.

**Data**

Yi and Chan (2010) analyzed the LCSH structure and reported that the LCSH structure offers a few large groups of highly interconnected main headings, but most main headings are connected to few. They reported that the most connected groups were associated with the domain of Science; consequently, this domain is chosen for this study since it offers the most complex structure.
The specific data sets are comprised of 204,430 LCSH topical authority records provided by the McGill University Libraries in January 2008, and 130,940 bibliographic records housed in the Schulich Science and Engineering library. As this collection stems from a sizable academic library, we have no reason to believe the results of this study would not be relevant to any LCSH organized science collection of equivalent size found throughout the world.

**Extracting LCSH Strings from the Bibliographic Records**

The vast majority of bibliographic records (93.47%) are indexed by at least one topical LCSH string. These contain 64,681 unique topical LCSH strings, and each record contains an average of 1.94 LCSH strings. The distribution of the number of bibliographic records per LCSH string is shown in Fig. 1 (Error! Reference source not found.). It shows a power law with a long tail containing 64.86% of LCSH strings that are assigned to a single bibliographic record. The top 1% most assigned LCSH strings account for 30% of all assignments, and this small group of LCSH strings provides access to 43% of bibliographic records.

The systematic cataloging practice of creating subdivisions to assign most specific LCSH string (i.e., law of specificity) enhances the long tail of quasi-empty LCSH strings. This hampers the grouping function of subject indexing, and it is a usability issue since, in the majority of cases, users would encounter LCSH strings that contain very few or often a single bibliographic record.

**Fig. 1: Distribution of bibliographic assignments per LCSH strings**

(64,681 LCSH strings, random sample of strings shown)
Matching with Authority

Table 1 describes the process of matching the LCSH strings extracted from the collection with the authority records. It shows that only 23% of the LCSH strings assigned to the collection exactly match an established LCSH string from the authority records. The systematic cataloguing practice of assigning subdivisions explains why an additional 75.7% of LCSH strings were matched with an authority record only after progressively removing at least one of its subdivisions. As a result, 98.7% of collection LCSH strings are matched to 18,318 authority records. This group of authority records is 71.3% smaller than the number of LCSH strings assigned to the collection, yet it provides access to over 99.8% of the collection. This shows that the collection LCSH strings can be adequately matched to the authority records.

**Table 1: Process of matching collection LCSH strings with the authority records**

<table>
<thead>
<tr>
<th>Authority records</th>
<th>204,429</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bibliographic Records</td>
<td>122,393</td>
</tr>
<tr>
<td>LCSH strings in collection</td>
<td>64,681</td>
</tr>
<tr>
<td>Matching LCSH strings with authority records</td>
<td></td>
</tr>
<tr>
<td>Exact string matches</td>
<td>14,903</td>
</tr>
<tr>
<td>Subdivision matches</td>
<td>48,960</td>
</tr>
<tr>
<td><strong>Total Matches:</strong></td>
<td>63,863</td>
</tr>
<tr>
<td>Impact of matching LCSH strings to authority records</td>
<td></td>
</tr>
<tr>
<td>Matching authority records</td>
<td>18,318</td>
</tr>
<tr>
<td>Collection Access</td>
<td>122,202</td>
</tr>
</tbody>
</table>

A visual inspection of the unmatched LCSH strings showed that many have slight spelling or format differences with an established string, or they are written in foreign language (e.g., French). Further analysis of these unmatched LCSH strings is beyond the scope of this research.

Matching collection LCSH strings with authority records accentuates the grouping function of subject indexing. The 71.3% reduction in the number of subjects representing the collection creates a 240.5% increase in the average number of bibliographic records per subject. Specifically, there are on average 3.7 bibliographic assignments per LCSH string, as compared to 12.5 per matched authority record. The distribution of the number of bibliographic records assigned per authority record still follows a power law: the top 1% most populous authority records comprise 27.26% of the total subject assignments and provide access to 41.82% of bibliographic records. The distribution’s long tail of subjects containing a single bibliographic record is decreased from 64.8% for the collection LCSH strings, down to 34.6% for the matching authority records.

Building the LCSH Hierarchy

The 18,318 established LCSH strings matched from the bibliographic records define a subset of the LCSH structure needed to browse the specific collection. Of these, a portion includes broader terms permitting the construction the LCSH hierarchy which represents the collection to be browsed.
The connectivity of an authority record is described in terms of its number of children and parents. Error! Reference source not found. describes the connectivity of the LCSH structure and its subset assigned to the science and engineering collection. The rows contain the count and percentage of orphans (i.e., LCSH without broader term). The first and second columns contain the percentage of orphans in the complete LCSH structure as compared with the subset of authority records assigned to the collection. They show that the science portion of LCSH has 45% fewer orphans (39.79 – 21.74) suggesting that science is a domain of LCSH with relatively fewer orphans. This would concur with Yi and Chan (2010) who reported that the domain of science was the most connected portion of the LCSH structure.

**Table 2**: LCSH hierarchy orphans, complete structure Vs. subset assigned to our science collection

<table>
<thead>
<tr>
<th></th>
<th>All LCSH (204,429 strings)</th>
<th>Science subset of LCSH assigned to the collection (18,318 strings)</th>
<th>All LCSH Orphans matched</th>
<th>Science subset Orphans matched</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Orphans</td>
<td>81,357</td>
<td>3,983</td>
<td>22,012</td>
<td>598</td>
</tr>
<tr>
<td>% Orphans</td>
<td>39.79</td>
<td>21.74</td>
<td>10.77</td>
<td>3.26</td>
</tr>
</tbody>
</table>

A large number of independent branches (i.e., orphans) is a usability issue for browsing; however, a visual inspection of orphan and island LCSH strings reveals that most are established subdivisions. For example, *Aircraft accidents--Human factors* is an established LCSH string whose authority record does not specify a broader term. Intuitively, *Aircraft accidents--Human factors* could be a narrower term to *Aircraft accidents*; indeed, users might find it difficult to understand why it is not.

**Matching Orphans**

The number of orphan LCSH strings is mitigated by automatically adding a broader term relation between established subdivided LCSH strings and their respective main heading. In this case, Aircraft accidents--Human factors receives a broader term relation with Aircraft accidents. Other orphan strings suggest they have a comma delimited broader term. For example, Abbreviations, English is an orphan branch although Abbreviations is a main heading. An automatic routine connects these types of branches as narrower terms with their main heading located left of the comma. Although not explicitly defined by LC, these two types of implied broader relations are conceptually sound for the purpose of creating a usable LCSH browsing tool.

Error! Reference source not found., 3rd and 4th columns present the impact of the orphan matching operations. Orphans are reduced by 72% (39.79 – 10.77) for the complete LCSH structure, and by 85% (21.74 – 3.26) for the subset of authority records assigned to the collection.

Some of the remaining orphans assigned to the collection are roots of highly general established LCSH strings such as Science, or Civilization; however, many are clearly not. For example, the established LCSH string Aluminum compounds has no broader term—users might wonder why this concept is not a narrower term of *Aluminum*. Attempts to exploit this pattern were abandoned since it was difficult to know which subset of words found in the string should constitute its broader term(s) and false relations could be created. For example, *Fire management* has no broader term but it could arguably belong to the domain of forestry; however, it should not be a narrower term to *management*. As a result, these types of orphans were left unaltered as root nodes.
Collection Specific LCSH Hierarchy

The LCSH hierarchy representing the collection is comprised of a total of 18,318 authority records. Beyond the number of unique established strings, the LCSH structure can be described in terms of numbers of narrower/broader terms. We wish to compare the following distributions:

- Numbers of children per authority record
- Numbers of children per authority record assigned to the collection
- Numbers of parents per authority record
- Numbers of parents per authority record assigned to the collection

Fig. 1 contains two panels showing the box-and-whisker diagrams representing the shapes of these distributions at different scales. The top panel shows the true range whose size forces the downscaling of the boxes to a single horizontal line. To facilitate box comparisons, the bottom panel shows the log of the range.

Starting from the left of Figure 1, bottom panel, the top of a box marks the 3rd quartile showing that at least 75% of authority records have one or no children. In comparison, the 2nd box represents the distribution of children per authority record assigned to the collection. It shows that at over 50% of assigned authority records have at least one narrower term, and that the average number of children (multiplication sign \(\times\) in Figure 1) is over four times higher than for the complete LCSH structure.

On the top panel, the whiskers show an impressive range of up to 586 children for a single authority record. These broad ranges suggest that the LCSH hierarchy is highly unbalanced. In the majority of cases, users are offered a single narrower term; however, in a few cases they are asked to choose from dozens or even hundreds of narrower terms. This could be a usability issue for browsing, and balancing the LCSH hierarchy will be the subject of future research.

Fig. 1: Box-and-whiskers diagrams showing the distributions of the number of children and parents per authority record, for the complete LCSH structure and the subset assigned to the collection. Average as multiplication sign \(\times\).

---

Note that some of these relations are the result of the orphan matching process. As a result, this distribution includes relations beyond those specified by LC.
The distributions of the numbers of parents per authority record show that 75% of authority records have one or no broader terms, as compared with 50% for the assigned LCSHs (see figure 1, 3rd and 4th boxes). Browsing such a lattice (Besag, 1974) could be a usability issue since traditional hierarchy browsing tools (e.g., file explorer) accept a single parent per child, and duplicating a branch under each of its parents may create a highly redundant structure with multiple copies of an authority record and all its descendants. Transforming the LCSH lattice into a valid tree structure and evaluating the impact on redundancy will be the subject of future research.

Recommendations

Based on the results of this study, we can provide some guidelines for those wishing to develop a LCSH based collection browsing tool. Firstly, matching the majority of LCSH strings assigned to the collection with an established string from LC can only be done by ignoring subdivisions; however, this is not a systematic rule since there are established subdivided strings. The algorithm should start by attempting to match the complete string. If this is unsuccessful, it should remove a single subdivision, attempt to match, and repeat until the string contains no further subdivisions. For our collection, this approach matched 98.7% of LCSH strings found in the bibliographic records.

Constructing the LCSH hierarchy is hampered by subdivided authority record strings which do not specify a broader term. Yi and Chan (2008) chose to ignore subdivided strings because they felt that associating a subdivided string with its main heading would not reflect a true broader relation (Yi & Chan, 2009). Our position is that the LCSH structure should have broader relations between main headings and their subdivisions; this is deemed conceptually sound for the purpose of browsing the LCSH structure. For our collection, these types of implied broader relations reduced the number of orphans by 72% for all authority records, and 85% for the subset assigned to the collection. This may facilitate browsing since users no longer have to choose from thousands of independent top level subjects.

Conclusion

The objective of this study was to construct and characterize the LCSH hierarchy containing a specific collection to be browsed. The analysis of a science and engineering collection indexed using LCSH showed that matching the LCSH strings assigned to bibliographic records with an established string from LC produced a 71% reduction in the number LCSH strings containing the collection. This operation also accentuated the grouping objective of subject indexing with a 30% reduction in the number of LCSH strings assigned to a single item. These findings are promising for the development of a smaller LCSH hierarchy with more of its LCSH strings providing access to significant groups of bibliographic record (as opposed to a single record).

We have also confirmed the presence of a power law distribution of LCSH string assignments; a small group of subjects is assigned to large numbers of bibliographic records, most subjects are assigned to few or often a single record. This finding provides a promising avenue to further reduce the size of the LCSH browsing structure based on information density. Users could be given features to dynamically reduce the size of the LCSH hierarchy by revealing the few major LCSH branches which contain most of the collection. This will be the subject of a future research.

References


Pierre Tirilly – University of Wisconsin, Milwaukee, USA
Charles-Antoine Julien – McGill University, Canada

Random Walks for Subject Hierarchy Simplification

Abstract
Although subject hierarchies are widely used to index document collections, few tools leverage their structure to facilitate collection browsing. This is mostly due to the complexity of such structures that include thousands of nodes. This paper proposes a new approach to simplify subject hierarchies based on the distribution of documents among the nodes. A random walk algorithm simulates the route of a user within the hierarchy, under the assumption that the user is attracted by the most populated nodes. Poorly visited nodes can be identified and eliminated, leaving a structure containing only the nodes that best represent the content of the collection. Experiments on a collection indexed using the Library of Congress Subject Headings (LCSH) show that, as compared to the state-of-the-art simplification method, the random walk-based approach gives access to a larger part of the collection for the same structure size, and offers more flexibility to customize the complexity of the structure.

Introduction
Using a controlled vocabulary is a classical way to index information collections, either physical ones (e.g. library collections organized using the Library of Congress subject headings LCSH) or digital ones (e.g. medical articles on PubMed organized using the Medical subject headings MeSH). Users can retrieve relevant information within an indexed collection by querying the system with terms from the controlled vocabulary using a traditional search box interface (Lancaster, 1986). However, when the user is a novice of the domain covered by the collection and is not familiar with the indexing vocabulary (Belkin et al, 1982), or when the problem is ill-defined (Marchionini & Shneiderman, 1988), browsing the collection is more appropriate than performing direct searches.

Browsing the collection can be supported by the implicit structure of the controlled vocabulary. Indexing terms are usually organized using different types of relationships, especially broader/narrower term relationships that indicate that one term is more general or specific than another. Documents indexed by connected terms are thematically close and should be displayed next to each other in a browsing interface. The relationships between the vocabulary terms define a global hierarchical structure of the vocabulary, and by extension of the indexed collection, from the most general subjects to the most specific ones. Formally, this structure is a directed acyclic graph, where the nodes correspond to indexing terms (and their associated documents) and the edges correspond to the relationships between terms. Figure 1 shows an example of such a graph. This graph can be explicitly provided with the indexing vocabulary (as in the case of MeSH 1), or can be reconstructed using the term relationships provided in the authority records (as in the case of LCSH (Julien, 2010)). This graph could serve as a basis for effective browsing interfaces.

Current browsing interfaces generally do not leverage this structure, preferring alphabetical lists of indexing terms, or sequential browsing of the terms that hide the global structure of the collection (Larson, 1992; Richmond, 1959). Some interfaces (Papadakis et. al., 2009; Yi & Chan, 2008) propose to explore the global structure of the vocabulary, but independently of any indexed collection, making them poorly suited to information retrieval tasks.

1http://www.nlm.nih.gov/mesh/
One major drawback of using the vocabulary structure for browsing is the daunting complexity of controlled vocabularies, which are typically composed of thousands of terms. This prevents from displaying the general structure of the collection, allowing showing only local portion of it. This problem can be tackled by simplifying the structure to display only the subjects that correspond to the major topics of the collection. Such a simplification is possible as the distribution of the documents among subjects follows generally a power law, i.e. most documents are assigned to a limited number of subjects, and most subjects are assigned a very small amount of documents (Smiraglia, 2001). Julien et al. (2012) proposed an ad-hoc simplification algorithm based on this property and successfully applied it to a collection indexed by LCSH.

In this paper, a new algorithm to simplify subject hierarchies is proposed. This algorithm is based on random walks on the vocabulary graph, a technique that simulates the browsing behavior of the users to extract the subjects that are most likely to be visited by the user. Experiments on a collection of documents indexed using LCSH show that this approach produces a drastic reduction in the size of the hierarchy while providing access to most of the documents of the collection. Compared to the existing simplification method, random walks allow fine tuning of the model to fit the behavior of the user.

Method

Random Walk on Subject Hierarchies

Random walks (Spitzer, 1976) are a family of simulation techniques with applications in physics, chemistry and computer science. Random walks on graphs (Lovasz, 1994) are a specific case for graph analysis, used for instance for network analysis or information retrieval. A random walk in a graph simulates the walk of an agent through the graph. It determines the probability that the agent ends up on each node of the graph, according to its starting point, the transition probability to walk from one node to one of its neighbors, and the structure of the graph. The final probability distribution on the graph nodes can be computed exactly or efficiently approximated using an iterative algorithm.
In a subject hierarchy, random walks can be used to simulate the behavior of users browsing the structure of the hierarchy to find information. As illustrated in Figure 2, the user starts from a subject (here, node N3) with starting probability (\(Pr(N3)\)), then moves to a broader or narrower subject (here, parent N2) with a given transition probability (\(Pr(N2|N3)\)), then moves to a new subject (child N4), etc. The specific behavior of the user depends on the way the probabilities are computed.

We model the random walk based on the following assumption: the user will tend to visit the most popular subjects of the hierarchy. The popularity of a subject is considered to be proportional to the number of documents accessible from this subject. To do so, the starting probability of each node is computed based on the proportion of documents that can be accessed from this node. This corresponds to a user who starts browsing the collection after a keyword-based search, assuming that users are more likely to query popular subjects. The transition probability from one node to another is computed based on the proportion of documents that can be accessed from the current node and the nodes connected to it. Two cases have to be considered for the transition probabilities: the case when the user goes down from the current node to a child node, i.e. a narrower subject, and the case when the user goes up to a parent node, i.e. a broader subject. These cases correspond to different behaviors: browsing more general or more specific subjects. The relative importance given to each behavior can be controlled in the probabilistic model by a single parameter: the parent weight \(W\). The higher \(W\), the more the user tends to go up in the hierarchy. \(W\) is the main parameter of this random walk model and is included in the calculation of the transition probabilities.

**Random Walk-based Hierarchy Simplification**

In this paper, the node probabilities provided by the random walk algorithm are used as a basis to simplify the hierarchy to allow users to browse the information collection easily. The probabilities of node visits provided by the random walk are used to determine the significant nodes of the graph, i.e. the nodes that are very likely to be visited by the users, based on their connectivity and the number of documents to which they give access. Significant nodes are nodes whose probabilities of being visited are above a given threshold.

Fig. 2: Example of random walk in a subject hierarchy. Red lines show the path followed by the user and the associated probabilities. Grey nodes are not visited
The simplification procedure is based on the two simplification operations proposed in (Julien et al., 2012): pruning and compression. Pruning (illustrated in Figure 3) corresponds to the elimination of the least significant leaf nodes \(^2\) of the hierarchy, i.e. the leaf nodes whose probability of being visited is below a given threshold. The non significant leaf nodes are first pruned, leaving their parents as new leaf nodes. This process is repeated until only significant leaf nodes remain in the hierarchy, i.e. nodes whose probability of visit is above the pruning threshold. Compression (illustrated in Figure 4) corresponds to the elimination of non significant non-leaf nodes, i.e. nodes with children whose probability of being visited is below a given compression threshold. When such a node is eliminated, the global consistency of the hierarchy is maintained by adding all the children of the removed node as children of its parents. Pruning and compression are applied once each successively to obtain the simplified hierarchy. The simplification process adds two parameters to the complete algorithm: the pruning threshold and the compression threshold.

**Fig. 3: Pruning process in several steps.** Node N4 becomes a leaf node after the first step and is eliminated at the second step. Grey nodes denote non significant nodes.

**Experiments**

**Data**

The dataset contains 204,430 LCSH authority records provided by the McGill University Libraries in January 2008, and 130,940 bibliographic records from the Schulich Science and Engineering library. The portion of LCSH that covers the bibliographic records is extracted and the directed acyclic graph representing the subject hierarchy is constructed using the methodology proposed in (Julien, 2010).

**Simplification Performance and Behavior**

Figure 4 presents the proportion of documents that is accessible in the simplified hierarchy with respect to the size (number of nodes) in the hierarchy, using either our

---

\(^2\) A leaf node is a node with no children. It corresponds to subjects of the hierarchy with no narrower subject associated.
simplification method or the baseline method proposed by Julien et al. (2012). The different points are obtained by using different values of the parameters of the algorithms. For the random walk-based simplification, the ranges of the parameters are: parent weight $W$ in $[0;0.9]$ with steps of 0.1, and compression and pruning thresholds in $[10^{-1},10^{-1}]$ with steps of one order of magnitude. For the baseline method, the parameters and results are those reported in (Julien et al., 2012).

Compared to the baseline, the random walk approach can provide a better access to the collection for an equal number of nodes in the hierarchy. It better isolates the nodes that give access to most of the collection. Moreover, the parameters of the random walk seem to allow tuning the simplified more finely than the baseline approach, providing a better control on the resulting hierarchy. However, it is necessary to choose the values of the parameters of the random walk-based simplification carefully, as some sets of values may result in a degradation of the performance. From the analysis of the performance of the system with respect to the values of the parameters (not detailed here), several empirical rules can be identified. Optimal results are obtained with values of $W$ in the range $[0.1;0.6]$, and for pair-wise values of the pruning and compression thresholds that are similar (difference up to two orders of magnitude). Within these ranges of values, the two following rules hold. First, the higher $W$, the more the hierarchy is simplified. Second, $W$ allows a finer tuning of the size of the hierarchy than the pruning threshold, and the compression threshold allow a finer tuning than the two others. In other terms, modifying the pruning threshold has more impact on the structure (number of nodes and accessible documents) than modifying $W$, and modifying $W$ has more impact than modifying the compression threshold.

**Fig. 5:** Proportion of accessible documents with respect to the number of nodes (subject headings) in the simplified hierarchy. The different points are obtained using different values of the parameters of each algorithm.

---

3 This point has to be checked with a more systematic testing of the parameters of the baseline than what is provided in (Julien et al., 2012).
**Shape of the Simplified Graphs**

Table 1 presents some statistics obtained from the original LCSH graph, three graphs obtained using the baseline method and three graphs obtained during the random walk (RW) method. As already shown in the previous section, the random walk method offers better rates of collection access for similar numbers of nodes. In terms of depth (lengths of paths between two nodes) and breadth (number of children per node), the graphs obtained through a random walk-based simplification are closer to the original graph, whereas the baseline approach reduces more the depth and the standard deviation of the breadth. One explanation to this behavior lies in the design of the two approaches. The baseline algorithm makes local pruning/compression decisions based on the number of accessible documents only, and tends to eliminate long paths and nodes with numerous siblings. The random walk assigns probabilities to nodes based both on the number of accessible documents and the connectivity of the nodes (the odds to visit a highly connected node and its surroundings are higher than for isolated nodes). As a result, the shape of the resulting graph is closer to the shape of the original graph, as the latter is taken into account when walks through the graph are simulated. As stated by Julien et al. (2012), knowing what shape of graph offers better browsing capabilities to the user is an open issue. Usability studies are necessary to decide which type of graph, i.e. which algorithm would be preferable.

![Table 1: Properties of a sample of simplified graphs](SD stands for standard deviation).](image)

**Conclusion**

This paper presented a new method to simplify subject hierarchies based on the distribution of the documents that it indexes. By simulating the “walk” of users within the hierarchy, the algorithm is able to identify the nodes that provide access to the majority of the document collection. Compared to the existing simplification methods, this approach provides a better access to the collection for hierarchies of equal sizes. Some empirical guidelines to tune the size of the resulting hierarchy were provided.

Future work includes a more extensive evaluation of simplified subject hierarchies involving actual users to determine which properties of the structure are preferable to visualize and browse collections (connectivity, depth...). Moreover, the random walk model presented here could be further modified and improved to include different behaviors of the users. For instance, using random walks with restarts would allow simulating the use of horizontal relations between subjects (related terms in LCSH) or the behavior of users starting over their search from a new node.

**Acknowledgments**

This work received support from the research group for information retrieval (RGIR) of the University of Wisconsin-Milwaukee.
References


Faceted Taxonomy as a Mechanism for Browsing and Accessing Digital Libraries of theses and Dissertations: A Case Study

Abstract
We present the results of research that sought to prioritize the user’s information needs in relation to digital libraries of theses and dissertations. To this end, a faceted browsable taxonomy was created to offer the user a search option that is easy to use and interpret, with organized information. In the development of the study, we applied: the socio-cognitive approach of domain analysis theory, Ranganathan’s facet analysis, and content analysis, with the application of the thematic categorical analysis technique. We concluded that the browsing mechanism facilitated the visualization, exploration and retrieval of detailed information on research results by the academic community.

Introduction
Scientific communication is essential for development of human knowledge, since new knowledge emerges from something already discovered. It could be argued that the acquisition of new knowledge is facilitated when information on different theoretical, epistemological and practical approaches are easily and readily accessed. In this context, one of the principal functions of information retrieval systems (IRSS) is to promote and facilitate the sharing and transfer of information. IRSS have been developed to cater to the needs and demands of specific user communities. Among these are digital libraries of theses and dissertations (DLTDs), which have grown in number and coverage since the start of the 21st Century (Rosetto, 2008). The DLTDs are considered a type of specialized information service, as they are usually designed to satisfy a particular user. In Brazil, the digital availability of scientific documents, especially of theses and dissertations, is mandatory. In 2006, this was made binding by Regulation 13/2006, created by the Coordination for the Improvement of Higher Level Personnel (CAPES), the body responsible for the evaluation of graduate programs in Brazil. Despite this enforcement, there are still different problems in effectively accessing these documents for a variety of reasons reported in literature; titles that are not greatly representative of research content, assigning insufficient keywords or even incomplete abstracts (Borko & Chatman, 1963; Vieira, 1988; Nogueira, 1997; Gil Leiva, Fujita & Rubi, 2008; Eliel, 2008; Souza, 2009).

In view of these problems, a faceted browsing taxonomy was created for a DLTD, with the aim of offering structured (organized) information and with results already refined, based on the identification of user needs. The taxonomy enabled information to be organized, providing consistency and standardization of the information/content of theses and dissertations, and the facility for browsing through the terms of the faceted structure to facilitate retrieval of document content. The design proposed offers the user a search interface with access to a detailed information resource.

The Development of a Faceted Browsing Taxonomy
Taxonomy, in its traditional sense, presents a domain in the form of a simple hierarchy. A faceted taxonomy is capable of representing a domain in facets and identifying relations
between them, allowing you to establish the multi-dimensionality of a term in this domain. In the digital environment, the browsing possibility is based on the indexing and association of ideas and concepts, organized in the form of links, which act as virtual doors (gateways) that open up paths to other information” (Bonilla, 2005, p.139). This “association of ideas” is the very sense of faceted browsing that facilitates the inter-relation and retrieval of information, an approach called “faceted browsing paradigm” (Broughton, 2006; Tzitzikas & Analyti, 2007; Sacco & Tzitzikas, 2009). Besides this, faceted browsing allows the user to formulate the search progressively, ensuring its refinement and thus eliminating any possibility of “empty results” (Tunkelang, 2009).

A faceted browsing taxonomy is a system of categories that share a single logical criterion of subdivision, organized in the form of browsable links, in which there is a hierarchy of facets and sub-facets under each category, assigning different dimensions to an object (Maculan, 2011). This system is characterized as hypertextual and, according to Furtado (2010), is a controlled association of ideas based on a limited number of associations permitted by the system and imputed by the designer’s communication intention.

Following this assumption, we pre-establish a communicative purpose for the content of theses and dissertations in a DLTD. Consequentemente, criamos o mecanismo navegável, no qual um usuário pode iniciar sua busca em qualquer ponto da taxonomia facetada, de forma progressiva e dentre um conjunto limitado de itens e conexões. As a result, we create the browsable mechanism, in which the user may begin searching at any point in the faceted taxonomy, in a progressive manner and within a limited set of items and connections. Esse estudo de caso é apresentado a seguir. This case study is presented below.

The Case Study
This case study used the theses and dissertations digital library (DLTD) of the Federal University of Minas Gerais (UFMG). The investigation’s universe was the set of theses and dissertations submitted as part of the Graduate Program in Information Science (PPGCI) of the School of Information Science (ECI). A total of 290 documents - 62 theses and 228 dissertations originating from the PPGCI’s research group on „Organization and Use of Information” (OUI) between 1998 and 2009 were made available in July 2010 in the DLTD database. This study was limited to an analysis of the corpus that was formed of theses and dissertations a coming from OUI and the number totaled 41 documents - 12 theses and 29 dissertations.

The Methodology Applied in the Development of TAFNAVEGA
The development of faceted browsing taxonomy (TAFNAVEGA) began with the application of the domain analysis theory to identify the domain, the DLTD user and their informational demands. Next, the set of basic thematic classes (CBT) was determined, based on the facet analysis theory using Ranganathan principles. The CBT set was validated using the Content Analysis method, with the thematic categorical analysis technique (TCAT) based on Bardin (2009) and applied with Moraes” (1999) methodological proposal. This methodology is explained below.

Theory of Domain Analysis
As Jesse H. Shera, who had already thought of information as originating and dependent on the social and cultural context, at the beginning of the 1970s, Hjorland and Albrechtsen (1995) also affirm that knowledge is a historical, cultural and socially developed product. Hjorland’s, a pioneer in the formulation of the domain analysis theory, approach was seen as the most appropriate for domain identification. To describe the investigated domain, we
used the pragmatic method suggested by Hjorland (2002), based on the determination of objectives, goals and usefulness of the knowledge generated in this domain. In this sense, we verified the information flow at the heart of the discourse community, identifying a practical sense of production, communication and use (reception) for the information content of theses and dissertations. To obtain these elements, we answered the questions: „who produced the information”, „in what social context was this information produced”, „what was the purpose of the production” and „who uses the information resource”. In this process, we applied a procedure in four stages: (1) domain identification, in which the domain was characterized from the following inquiries: what is the domain?; Is the domain known?; What type of information does this domain produce?; (2) data analysis, with the setting of relations and functions between the information, based on the following questions: „who produces the information identified in the previous stage”?; „For what purpose is the information produced?”; „who is interested in using the information produced?”; (3) domain modeling, in which the information considered important for the information service to be created is selected, and is aligned to the objective proposed, based on the data collected in the previous analysis; (4) collection and selection of data and validation, in which specialized literature of the domain is consulted to support the previous stage, guided by the question: „who is the user that is interested in the content of these academic works, such as theses and dissertations?” Following this methodology, we achieved the results shown below.

Table 1: Identification of domain and information needs

<table>
<thead>
<tr>
<th>Stages</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
<td>Scientific community from the IS area, mainly researchers from the Organization and Use of Information line of research.</td>
</tr>
<tr>
<td>Analysis</td>
<td>Each researcher needs to know what was already done to advance their knowledge; needs to widen the knowledge and enhance existing scientific methods and theories; needs to understand, expand and apply previously known scientific knowledge, without any reworking.</td>
</tr>
<tr>
<td>Modeling</td>
<td>Researcher’s information needs: a) themes that were already investigated; b) methods and theories applied in research; c) results and products achieved in research.</td>
</tr>
<tr>
<td>Collection, selection and validation</td>
<td>Literature from the area, especially research methodology, confirm information needs of researchers (modeling stage) and indicate that each researcher seeks the approval of their peers, as well as academic visibility. The information content of the theses and dissertations is produced by researchers and it is the researchers themselves who have a fundamental interest in this information content.</td>
</tr>
</tbody>
</table>

Based on these results, we began the identification of basic thematic classes (BTC) to represent the content of the theses and dissertations domain. For this, we applied the facet analysis theory, a principle that guides the development of faceted systems. There was the decomposition or defragmentation of documents in their constitutive parts, in search of their essential elements, from fundamental categories stipulated by Ranganathan: PMEST, i.e. Personality, Material, Energy, Space and Time. In this way, we continued with the facet analysis method in three stages: (1) selection of the documents to be analyzed; (2) reading of documents for the identification of subjects and their constitution, using dissection (segmentation of the domain in same level terms); (3) selection of more relevant concepts to compose the BTC set and determination of their ordering. At the end of this procedure, we achieved the result shown in Chart 2.
Table 2: The set of basic thematic classes

<table>
<thead>
<tr>
<th>Ranganathan</th>
<th>Domain: Basic Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personality</td>
<td>Theme</td>
</tr>
<tr>
<td></td>
<td>Historical/contextual foundation</td>
</tr>
<tr>
<td></td>
<td>Theoretical foundation</td>
</tr>
<tr>
<td></td>
<td>Product (final result)</td>
</tr>
<tr>
<td>Material</td>
<td>Data collection</td>
</tr>
<tr>
<td></td>
<td>Type of research</td>
</tr>
<tr>
<td></td>
<td>Object</td>
</tr>
<tr>
<td>Energy</td>
<td>Methods</td>
</tr>
<tr>
<td>Space</td>
<td>Environment</td>
</tr>
<tr>
<td>Time</td>
<td>There is no class (year of publication)</td>
</tr>
</tbody>
</table>

With this result (Chart 2), we began application of the categorical thematic analysis technique (CTA) for validation of the nine basic thematic classes (BTC).

Application of the categorical thematic analysis technique (CTA)

Moraes's methodological proposal (1999) for CTA requires the intended objective of the analysis to be established. Therefore, we set as objective to: “verify if the set of basic thematic classes is adequate to represent the textual structure of the theses and dissertations, so that it may be used in the TAFNAVEGA”. Next, the CTA procedures adopted were: a) preparation of information, with document analysis; b) unitization or transformation of content in units of content and record, with the codification and achievement of content units (CU) and record units (RU) from the textual structure of documents, complying with literature on research methodology and specific norms; c) categorization or classification of units in categories; 4) description, with the definition of each category and continuous validation of the category set; e) interpretation, with an analysis of the result achieved.

At the end of the validation, it was noted that all scientific research is based on theoretical inputs on the investigated theme, which needs to know the state of the art of the subject treated, as well as the methods and techniques that were already applied to the research problem.

Moreover, we verified that the set of basic thematic classes (CBT) is the “simplified representation” of all the textual content of thesis and dissertation documents. We saw that there was recognition of the logical sequence of the ideas presented, in order to organize them and classify them, based on their chapters, as well as their descriptive data. As such, the faceted browsing taxonomy (TAFNAVEGA) is now composed of nine validated basic thematic classes (CBT).

The methodological path also includes the development of an algorithm for the extraction of concepts from thesis and dissertations. The algorithm called “the thematic categorical matrix for academic work” is outlined below.

The Thematic Categorical Matrix for Academic Work

The algorithm is a sequence of tasks, with clear and well defined instructions, with the objective of extracting document concepts to feed the set of basic thematic classes (CBT). This algorithm is composed of three columns: 1) the first column is comprised of the CBT set; 2) the second column is formed of queries as suggested by NBR norm 12676:1992 and by PRECIS’s (Fujita & Rubi, 2006) indexation system analysis principles; 3) the third column is composed of the textual parts of the theses and dissertations in which answers to the queries in the second column can be found.
Table 3: Matrix for extraction of concepts in theses and dissertations

<table>
<thead>
<tr>
<th>Basic Thematic Classes</th>
<th>Query (Norm 12.676) And Precis</th>
<th>Part of Textual Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1. Theme</td>
<td>What is the subject of the document?</td>
<td>Abstract / Introduction</td>
</tr>
<tr>
<td>C2. Empirical Object</td>
<td>Which empirical object was used/analyzed in the study?</td>
<td>Abstract / Introduction / Methodology</td>
</tr>
<tr>
<td>C3. Environment</td>
<td>Are the theme, empirical object and/or action considered in the context of a specific place or environment?</td>
<td>Abstract / Introduction / Methodology</td>
</tr>
<tr>
<td>C4. Type Of Research</td>
<td>In view of the nature (basic, applied), the approach (quantitative, qualitative), the objectives (exploratory, descriptive, etc.) or the procedures (bibliographical, case study, research-action, survey, etc.), how can the research be defined?</td>
<td>Abstract / Methodology / Results / Discussion of Results</td>
</tr>
<tr>
<td>C5. Data Collection</td>
<td>What specific instruments (questionnaire, interviews, audio-visual recordings, document collection, etc.) were used to perform the action?</td>
<td>Abstract / Methodology / Results / Discussion of Results</td>
</tr>
<tr>
<td>C6. Methods</td>
<td>Which specific methods were used to perform the action (for example, techniques or methods for data treatment: statistical modeling, structural analysis, codification, content analysis, etc.)?</td>
<td>Abstract / Methodology / Results / Discussion of Results</td>
</tr>
<tr>
<td>C7. Theoretical Foundation</td>
<td>Was any specific theoretical approach used (theories, hypotheses, presumptions, etc.) because of the nature of the object to be researched and the objectives intended?</td>
<td>Abstract / Review of Literature / Foundation</td>
</tr>
<tr>
<td>C8. Contextual/ Historical Foundation</td>
<td>Which themes were treated and reviewed, from bibliographical research, to contextualize the theme researched in a profound and consistent manner?</td>
<td>Abstract / Review of Literature / Foundation</td>
</tr>
<tr>
<td>C9. Results</td>
<td>What points did the research achieve, in consideration of the objectives proposed? Was there a formulation or reformulation of theory, creation of a method or a product?</td>
<td>Abstract / Results / Discussion of Results / Conclusions</td>
</tr>
</tbody>
</table>

Source: adapted from Maculan (2011).

The algorithm (Table 3) was used on the corpus of the research for analysis of documents” informative abstracts. The objective of the matrix was to guide and systematize the conceptual analysis of theses and dissertations for extraction of concepts. The concepts selected populated the faceted browsing taxonomy structure (TAFNAVEGA). The results of this application are described below.

Results and Conclusions

The results show that the TAFNAVEGA mechanism facilitated the exploration, search and recovery of content from documents, providing access to detailed research data, such as theories, methods and data collection instruments, allowing for the combination of this different information.

At the end of the conceptual analysis procedure, a total of 168 different indexing terms (facets and sub-facets) were collected, which populated the nine basic thematic classes (CBT). From the CBT set, we chose five to represent an empirical analysis and demonstrate the type of information that can be retrieved: (1) Theme; (2) Object; (3) Type of research; (4) Data collection; (5) Methods (Table 4).
Table 4: Analysis of data collected in the research body

<table>
<thead>
<tr>
<th>CBT</th>
<th>Percentage</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theme</td>
<td>100%</td>
<td>- Two strands in the OUI line of research: organization of information (66%) and use studies (34%).</td>
</tr>
<tr>
<td>Empirical object</td>
<td>100%</td>
<td>- Enough heterogeneity;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Special interest in OUI line of research with organization and access</td>
</tr>
<tr>
<td></td>
<td></td>
<td>to medical information, using medical charts (8%).</td>
</tr>
<tr>
<td>Type of research</td>
<td>46%</td>
<td>- Qualitative and comparative research prevails, and “case studies”,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>making generalization of results difficult.</td>
</tr>
<tr>
<td>Data collection</td>
<td>39%</td>
<td>- Predominance of the use of interviews and questionnaires.</td>
</tr>
<tr>
<td>Methods</td>
<td>68%</td>
<td>- Recurrence of use of facet analysis and analytic-synthetic method,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>based on Ranganathan’s studies, apart from statistical methods for user studies.</td>
</tr>
</tbody>
</table>

Source: adapted from Maculan (2011).

The textual content of thesis and dissertations is semi-structured and is often not easily retrieved by the researcher. In this sense, using the matrix to index the theses and dissertations facilitated the retrieval of detailed information about the research results, making them accessible to the researcher.

We concluded that on implanting digital libraries of theses and dissertations, with the objective of displaying the knowledge accumulated in concluded research, making them available digitally is not enough; value addition must be made as this research showed. In this case study, we created the conditions for information to be accessed, facilitating its retrieval. The faceted browsing taxonomy (TAFNAVEGA) served as mechanism for the structuring of information contained in academic works, allowing exploration of the DLTD by browsing facets and sub-facets, facilitating the search and, consequently, enhancing information retrieval.

References


---

130
Ontology as Boundary Object

Abstract
A lack of semantic interoperability in the multidisciplinary delivery of health care leads to poor health outcomes. This paper describes research that has lead to the development of an ontology based on SNOMED CT®. The ontology functions as a boundary object to bridge the semantic interoperability gap between members of multidisciplinary health care teams caring for patients with chronic diseases. Overall, there was strong agreement among the clinicians on the usefulness of the boundary object.

Introduction
The sharing of representations for cooperation between information systems or between humans and information systems requires “interoperability”, i.e., the ability of one system to communicate and interact with another system. For semantic interoperability, the data in the messages have meaning only when they are interpreted in terms of the subject domain of the system. However, users do not always know the model of the subject domain of a system. Semantic interoperability problems arise when user subject domain models differ from the system subject domain model. (Pokraev, 2005).

Among humans, the meaning of any set of terms, and the significance and utility of any taxonomy, can be evaluated only in the context of a community whose members are involved in similar activities and share similar values. (Wenger, 2002) In short, semantic interoperability among humans is tied directly to communities of practice, and to the negotiation of meaning that occurs within them. (Friesen, 1999) The question for semantic interoperability is how to communicate across these communities of practice.

One solution to the semantic interoperability problem is the introduction of boundary objects that serve as an interface between different communities of practice. (Bowker and Star, 1999) Boundary objects are shared by different communities but viewed or used differently by each of them and contain sufficient detail to be understandable to both parties, yet neither party is required to understand the full context of use by the other. (Denim, 2003)

An example of a boundary object is an electronic health record. It is used by doctors, nurses, hospital administrators, insurers, government, etc. Each community of users will have a slightly different view of and understanding of the health record, but the health record serves as an interface among these communities.

In this research, we created and evaluated a boundary object to interface among different communities of practice in the treatment of chronic diseases that require multidisciplinary care teams. These multidisciplinary care teams may consist of doctors, nurses, physiotherapists, dieticians, psychologists, etc. Each of these disciplines represents a community of practice that must work effectively with the other disciplines (communities) in order to deliver appropriate health care. The approach taken in this research was to develop an ontology, based on the SNOMED CT® vocabulary that encompassed the specialized vocabulary of each of the disciplines. The resulting ontology serves as a boundary object that maps from one specialized vocabulary to the others.

The Need for Semantic Interoperability in Multi-disciplinary Care
Multiple Chemical Sensitivity (MCS) and chronic pain are examples of complex chronic medical conditions for which there is growing evidence of the benefits of multidisciplinary
care management (Clark, 2000; Fox et al., 2007). Patients with these conditions exhibit a wide range of symptoms from physical to psychological (Dysvik et al., 2005; Kolstand et al., 2006). The nature of the multifaceted symptoms requires a multidisciplinary management approach with a comprehensive assessment of all factors affecting the individual’s health such as physical, psychosocial, economic, nutrition and medical areas of health focus. The increasing occurrence in primary care, of medical errors, repeated medical tests and poor understanding of the health conditions has made it critical to improve communication among clinicians involved in the care management of such diseases (Dobscha et al., 2009).

**Boundary Objects to Address Communication Gaps**

The intention of using a boundary object approach was to enhance collaboration among multidisciplinary care providers across communities of practice in the heterogeneous domain of complex chronic conditions (Star and Griesemer, 1989). This research has considered the heterogeneous domains of two complex chronic conditions, Multiple Chemical Sensitivity (Fox et al., 2007) and chronic pain (Clark, 2000) to investigate, develop and evaluate the boundary objects approach to improving collaboration. The characteristics of the boundary objects developed in the study had to fulfil certain requirements in order to overcome the challenges of heterogeneity in the domain knowledge. Using the boundary objects approach, we have developed a pragmatic level of interoperability as outlined by Carlile (2004) for clinicians collaborating in the care management of complex health conditions. We have also applied standardized forms of boundary objects taking into consideration the issues raised by Fujimura (1992) around too much flexibility leading to loss of identity. We have also designed the boundary objects such that they have the capacity to maintain a dynamic nature to enable the growth of new knowledge (Gal et al., 2005).

**SNOMED CT®**

SNOMED CT® or Systematized Nomenclature of Medicine SNOMED CT® (2005) is a comprehensive, multilingual, controlled clinical reference terminology, with comprehensive coverage of diseases, clinical findings, etiologies, procedures, living organisms, and outcomes used for recording clinical data. It is a relational, concept-based system with more than 300,000 unique concepts and more than 900,000 descriptions. Concepts are organized by defined relationships and organized in 19 hierarchies or facets. The main advantages of SNOMED CT® are that it is multiaxial, hierarchical and has the provision to express the underlying knowledge. In addition to clinical terms, SNOMED CT® includes general concepts such as occupation, social concept, physical object, life style, physical force, etc., which can be captured as well.

SNOMED CT® can be characterized as a multilingual thesaurus with an ontological foundation. Thesaurus-like features include concept–term relations such as the synonyms. It is a class hierarchy with extensive overlap of classes. The superclass (IS_A) Relation relates classes in terms of inclusion of their members. (Wikipedia, 2010) For example, the concept *Fibromyalgia* IS_A “disorder of skeletal muscle” IS_A “Disease”. A concept or class may have multiple parents. For example, the class *Vulvodynia* has two parents (*Disorder of Vulva* and *Disorder characterized by pain*).

SNOMED CT®s relational statements are basically triplets of the form Concept<sub>1</sub> - Relation<sub>x</sub> - Concept<sub>2</sub>, with Relation<sub>x</sub> being from a small number of relation types (called linkage concepts), e.g. *finding site, due to*, etc. The interpretation of these triplets is (implicitly) based on the semantics of a simple Description logic, which lends itself to the OWL language for ontology representation. (Wikipedia, 2011)
The structure of SNOMED CT® lends itself to the creation of a boundary object among multiple communities of practice through its class hierarchy with overlapping classes and its thesaurus-like features.

**Standardizing the Medical Chart Terminology**

SNOMED CT® was used as the reference terminology to standardize the terms retrieved in the chart audit process and in the representation of concepts in the ontology. One hundred patient charts were audited and the standardization of the chart audit concepts included the following steps.

- A clinical term was identified for standardization from the chart audit process when the term is a recurring term used to describe a patient profile.
- A manual search for an identical match to the source term was made using a SNOMED CT® browser.
- A search was made for alternative terms / synonyms when an exact match is not found with the same clinical meaning (i.e. concept match) to the source term.
- Terms with no matches in SNOMED CT® were identified.
- The multidisciplinary team reviewed the standardized terms for accuracy and completeness.
- A controlled vocabulary with the relevant grouping of standardized concepts in the patient profile domain was created.

Table 1 shows examples of inconsistent chart terminology from various areas of health and their standardization using SNOMED CT®. The number of standardized terms in each field (community of practice) developed for the terms extracted from the Multiple Clinical Sensitivity charts, were 356 medical, 136 physical, 122 psychosocial, 118 rehabilitation, and 80 nutrition.

The multidisciplinary team of clinicians re-coded 3 prototypical patient charts using the new vocabulary. The clinicians used a web-based form to generate the profiles for patients using the controlled vocabulary. The domain experts tested the accuracy, completeness and relevancy of the standardized concepts included in the controlled vocabulary.

<table>
<thead>
<tr>
<th>Table 1: Example of Standardized Terminology for Multiple Chemical Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Terminologies in clinical notes</strong></td>
</tr>
<tr>
<td>Fatigue, low energy, very tired, extremely tired, heavy feeling</td>
</tr>
<tr>
<td>Light sensitivity, hypersensitivity to light, intolerance to light</td>
</tr>
<tr>
<td>Fibromyalgia, FM, Myalgia</td>
</tr>
<tr>
<td>Poor balance, balance impairment, loss of balance, unsteady</td>
</tr>
<tr>
<td>Poor sleep, sleep problems, sleep issues, unrefreshed sleep, non-restorative sleep</td>
</tr>
<tr>
<td>BMI</td>
</tr>
<tr>
<td>SCL-90R</td>
</tr>
</tbody>
</table>

**Creating the Ontology**

The second stage of the two-staged approach was the creation of an ontology in the heterogeneous domain. This consisted of 3 phases: Development, testing and evaluation. The development phase included the experts in the domain specifying and organizing the
knowledge in the domain. This phase primarily drew the knowledge from the controlled vocabulary. The key concepts were explicitly related by establishing relationships and attributes in the domain. Multidisciplinary interactions in the management of symptoms were specified in the ontology through relations and attributes. Multidisciplinary classes were also created in the ontology that showed the involvement of various grouping of clinicians in the management of a specific grouping of multidisciplinary symptoms for patients. The knowledge in this phase was derived from the patient charts and from the domain experts. Instances from one hundred patient profiles were populated in the ontology.

The testing phase included the clinicians browsing the profile ontology developed in this research to examine the concepts in the ontology, the relationships between concepts, concept attributes and the individuals populating the ontology. Following this was an evaluation phase that included feedback from the domain experts on the overall usefulness of the ontology in patient care with emphasis on usefulness from a health discipline perspective, from other health disciplines and the multidisciplinary nature of interactions captured in the ontology.

**The MCS Ontology**

The ontology was created in Protege 3.4.2. Figure 1 shows the top level classes developed for the ontology. The Patient Profile has five different subprofiles representing five different disciplines (communities of practice) involved in delivering of health care to a patient with MCS.

*Fig. 1: Top level of ontology.*

Figure 2 is a graphical representation of two communities of practice (Psychosocial and Rehabilitation) having the common concept, *Education and or school finding*. This is an example of the ontology acting as a boundary object between different communities of practice.
Evaluation

Seven domain experts were recruited to participate in the review of the ontology. Clinicians used the Google ontology browser to navigate through the ontology and were given an option to query the multidisciplinary interactions and the patient profile knowledge that exists in the populated instances of the ontology. Clinicians then used a survey questionnaire to provide feedback on the usefulness of the ontology. They provided feedback on the overall usefulness of the ontology, relevance of the ontology to their health discipline, usefulness of viewing information generated from other health disciplines and the usefulness of viewing multidisciplinary interactions.

A strong level of agreement was obtained among the domain experts (Fig. 3). The highest level of strong agreement was obtained in the overall usefulness category. The highest level of agreement was obtained on the usefulness of the multidisciplinary classes. A small amount of disagreement was obtained in the category of usefulness of the information from other health disciplines.

Fig. 3: Feedback on the MCS profile ontology
Figure 4 shows the evaluation breakdown by individual clinicians. Over half the clinicians showed a strong level of agreement on the overall usefulness of the ontology and most categories of the survey questionnaires. There were no specific trends by a health discipline or by a category in the survey questionnaire.

**Fig. 4: Feedback on the MCS profile ontology by clinician.**

---

**Conclusions**

This research has shown that an ontology, based on a controlled vocabulary, can effectively act as a boundary object among multiple disciplines of care (communities of practice). The methodology was created for Multiple Chemical Sensitivities and validated for Chronic Pain, both complex chronic diseases. Overall, there was strong agreement among the clinicians on the usefulness of the boundary object.

A long-term evaluation to determine the effect of such a boundary object on health outcomes is being planned.

**References**


Towards an Ontology for Mathematical Modeling with Application to Epidemiology

Abstract
Mathematical modelling is a field of applied mathematics with applications in other disciplines. The availability of a formal ontology and derived benefits, such as the possibility of conducting automated reasoning about the ontological classes of the domains, greatly reduce the barrier of entry in the field for non-experts, while helping the establishment of a more precise and controlled vocabulary among the domain experts involved in mathematical modelling. This work focuses on Mathematical Models applied to the natural sciences and as a case study the field of mathematical epidemiology has been chosen for this ontology. We propose the development of an ontology of mathematical models which is general enough and not restricted in its applicability, yet is developed considering the specific needs of a particular application domain.

Background and Purpose
Just as many other fields of science - or perhaps more than them - Mathematics needs abstract models to represent its core concepts and how they are related. Mathematical modeling is a field of applied mathematics which forms deep inroads into other disciplines. The availability of a formal ontology and derived benefits, such as the possibility of conducting automated reasoning about the ontological classes of the domain, greatly reduce the barrier for non-experts, while helping the establishment of a more precise and controlled vocabulary among the domain experts involved in mathematical modeling.

Mathematical Modeling is a vast domain, as it applies to almost every subject of human inquiry. To help delimit our endeavour, we decided to focus, for the purpose of this work, the scope of Mathematical Models to models applied to the natural sciences. This does not mean that a model which represents a purely abstract phenomenon cannot be adequately classified by this ontology, but rather that we will consider certain ontological categories, such as time and space, as fundamental components of the subset of models we are interested in. Furthermore, we have the field of mathematical epidemiology as the first target of this ontology. In summary, we propose the development of an ontology of mathematical models which is general enough as to not be limited in its applicability, yet is developed for the specific needs of a particular application domain.

A key motivation for this work is the possibility of using the resulting ontology as an aid to automate information extraction from the modeling scientific literature. Modelers build on previously published models which are modified to represent new application scenarios. Thus, to efficiently survey the literature for models of the type one is interested in, is of great importance. Moreover, finding values for a given model’s parameters can be an even harder task as it requires examining an even larger set of papers. Another important motivation was the perceived lack of ready-to-use ontologies at the level of abstraction desired. One of the oldest and more comprehensive ontologies in the domain of mathematics, EngMath (Gruber and Olsen, 1994), was never ported to a modern specification language such as OWL.

Other interesting initiatives we have identified include Ecolingua, an ontology for quantitative data in ecology (Brilhante, 2005), and PhysSys (Borst, P. et al., 1995), an ontology for physical systems. These ontologies address problem domains similar to ours and have served as a source of concepts, relationships, and the associated terminology. Closer to our scope is the concept of mathematical expression, included in the "Systems Biology Ontology" (Le Novère, 2006).
Another neighboring domain which has seen some recent formal ontological classification is that of algorithms used to simulate models (Juty et al., 2010), such as the ones we aim to describe in this ontology. We believe that KiSAO could naturally be used to extend ours in the specific domain of model implementation.

**Methodology**

Many authors agree that the approaches for building domain - and cross-domain - ontologies are, most of the time, specific and limited. One problem, from the methodological point of view is that there is no generally accepted pattern or set of phases for building ontologies (Fernández et al., 1999; Uschold and Gruninger, 1996). Despite the fact that great quantities of ontologies have already been developed by different communities as Chemistry (Gómez-Perez, Fernandez and Vicente, 1996) or Business Process Modelling (Gruninger and Fox, 1995), just to give a few examples – under different approaches and using different methods and techniques, there is no consensus about a „gold standard” methodology for the development process (Fernandez, Gómez-Perez and Juristo, 1997). The consequence is the absence of standardized activities, and these tasks are commonly driven in an artisan fashion and not as a scientific activity. Besides that, there is lack of a systematic explanation of how the theoretical approaches might be used pragmatically.

Rather than choosing a specific methodology, we have adopted the comparative approach formulated by Silva, Souza and Almeida (2008), that has analyzed the methods Cyc (Reed e Lenat, 2002), Tove (Fox, 1992), Enterprise Ontology Project (Uschold and King, 1995), Kactus (Bernaras, Laresgoiti and Corera, 1996), Methontology (Fernandez, M. et al, 1997), Sensus (Swartout et al., 1996) and 101 (Noy and McGuinness, 2001). This combined methodology, with slight modifications allowed us to choose the most well defined tasks for each of the ontology development phases. Following this hybrid approach, we have started working on the following tasks so far:

**Definition of the motivation scenario:** to support the task of classification and selection of mathematical models in the field of mathematical epidemiology;

**Definition of ontological scope:** given the cross domain characteristics, the ontology covers both mathematical concepts and communicable diseases epidemiology concepts, with special regard to Dengue fever. To keep it coherent with the boundaries of the domains, two separate ontologies were built and used together: One for mathematical models (MMO) and another for mathematical epidemiology (EMO);

**Integration:** After searching for existing ontologies in the general domain of our intended application, the following ontologies were imported and used: Protegè-Dublin Core Ontology in owl format ¹, Infectious Disease Ontology (IDO) ² and a publication ontology ³ to represent scientific publications;

**Conceptual Modeling:** The gathering of relevant concepts, terms and relationships was made mainly through the analysis of a corpus of scientific articles describing mathematical models in the field of Dengue epidemiology. The result of the harvesting of the literature was then filtered by the authors to select canonical concepts which formed the conceptual basis of mathematical epidemiology. Currently, both MMO and EMO contain 93 classes and 19 object properties. Together with the imported ontologies one can work with a total of more than a thousand classes.

¹ http://protege.stanford.edu/plugins/owl/dc/protege-dc.owl
² http://infectiousdiseaseontology.org/page/Main_Page
³ http://mapekus.fiit.stuba.sk/mapekus/ontologies/v0.2/publication
Implementation: the conceptual structures are implemented using Protégé 4.2.0 and OWL 2.0 as a formalization language/framework. The source codes and their documentations are maintained under version control on Bitbucket.

Maintenance: the maintenance of the ontology will be overseen by the authors along with current and future collaborators.

Results
After careful consideration of the existing ontologies in the more general domain of mathematics, we have constructed the first versions of MMO and EMO (Fig. 1&2). For the MMO, special attention was devoted in this first version, to concepts related to dynamic models. Dynamic models describe the temporal and/or spatial variation of biological or physical quantities. The conceptualization of space and time is of special importance, and is related to the type of model to be used (equations, automata, agents). Furthermore, MMO discriminate models regarding their representation into Mathematical (equations to be solved or calculated numerically) and Algorithmical (sets of logical rules to be simulated). This dichotomy is reified through the property “hasRepresentation” of a mathematical model. To further describe models, MMO defines classes such as “variable” and “parameter” which can be further described through properties such as: “hasContinuity”, to discern between continuous and discrete variables or parameters; “hasDomain”, to allow specification in which the entity takes its values; and “hasRandomness”, to separate fixed and random variables and parameters.

The MMO seeks to articulate concepts such as time and space. For example, when time is discrete and space is not explicitly considered, the mathematical model can assume the form of a difference equation. On the other hand, if time is continuous, differential equations are used. The MMO ontology uses properties to make these connections. For example “hasEquation some Difference_equation”.

The central class in the ontology is Mathematical_Model, which is shown in more detail in Figure 3. Six subclasses of models are contemplated so far. Models which are specific to a given domain should be defined on a separate domain ontology which imports MMO. For example EMO defines a class named „Susceptible_Infected_Recovered_Model” which is a subclass of „Epidemic_Model” which in turn is a subclass of „Dynamic_Model”.

EMO, the epidemic modeling ontology, covers the specific field of mathematical epidemiology. This discipline is built upon a set of well structured models, such as SIR (for directly transmissible diseases with full immunity, such as smallpox and measles), SIS (for those with no induced immunity, such as syphilis), SI (for those with no recovery, such as HIV), and so on (Keeling and Rohani, 2007). Such classical archetypes are expandable to include more host types (arthropod vectors or other animal reservoirs), multiple pathogens (four dengue serotypes, for example), multiple immunity classes (primary infections, secondary infections; vaccinated), creating a very diverse set of disease-specific models.

Disease specific models connect epidemiological concepts and mathematical representations. These entities, being from different domains, are not hierarchically related. In EMO, epidemiological concepts (hosts, vectors, transmission mechanisms, etc) are organized within the Mathematical Epidemiological Core Concepts class (Figure 2).

The scientific activity of creating disease-specific models based on more simple, archetypical ones has historically created several model lineages. One of the applications of EMO is to identify these lineages, which are not explicitly introduced in the ontology, but should emerge from the basic definitions. This is an ongoing process.

---

4 https://bitbucket.org/fccoelho/math-model-ontology
Fig. 1: The Mathematical Modeling Ontology
Discussion

The ontology presented in this paper aims to fill a void in the availability of formal ontology for the classification of mathematical models of natural systems. The literature on the subject “mathematical modeling” is vast and mature, which has allowed for a convergence towards a pragmatic taxonomy of model categories when it comes to the organization of educational materials. In this work, we have tried our best to incorporate this praxis, while adding the necessary rigor in the definition of the derived ontological classes. Despite our care, many aspects of the ontology are the inevitable result of our personal choice among alternative definitions. Whenever possible, references were added to the owl document to justify our classification.

This ontology is in progress, as we believe ontologies are “living artifacts”, which should evolve with the field it represents. The full ontology in owl format is available for download at https://bitbucket.org/fccoelho/math-model-ontology.
Fig. 3: Class Mathematical Model

References


Ontological Representation of Knowledge for Developing Information Services in Food Science and Technology

Abstract
Knowledge explosion in various fields during recent years has resulted in the creation of vast amounts of on-line scientific literature. Food Science & Technology (FST) is also an important subject domain where rapid developments are taking place due to diverse research and development activities. As a result, information storage and retrieval has become very complex and current information retrieval systems (IRs) are being challenged in terms of both adequate precision and response time. To overcome these limitations as well as to provide natural-language based effective retrieval, a suitable knowledge engineering framework needs to be applied to represent, share and discover information. Semantic web technologies provide mechanisms for creating knowledge bases, ontologies and rules for handling data that promise to improve the quality of information retrieval. Ontologies are the backbone of such knowledge systems. This paper presents a framework for semantic representation of a large repository of content in the domain of FST.

Introduction
With the rapid growth of on-line information, proper description of information is essential for effective information retrieval and filtering. Many proposals have emerged beginning with simple metadata and evolved to more complex and expressive structures. In the last decade, the exponential growth of information in Food Science & Technology (FST) research has brought to the forefront the need for automated knowledge acquisition, representation and analysis. Since the information is used by both computers and humans, there is a need for a knowledge representation that both can understand. The huge amount of information coupled with the need for automation also poses another problem, namely, heterogeneous sources of data are not compatible with each other making the task of finding and combining relevant information difficult. The most promising way to solve these problems is to use ontology-based knowledge representations. Ontologies provide common vocabularies using which knowledge can be represented in such a way that it is readable by both humans and intelligent agents. Such a framework can be seen as a model for representing organised knowledge in a given domain (e.g., FST). The idea of ontology is to describe and standardize basic concepts in the domain and support reasoning with axioms and predicates which can be used to define relationships among concepts. Such representations can be used for information integration from different sources by providing a common conceptual basis for the integrated information source.

Before presenting our framework, a description is given of the salient features of work considered most relevant to ontology-based information services in the FST domain. USDA National Nutrient Database for Standard Reference is a database made by the United Stated Department of Agriculture to be the major source of food composition data in the United States. Its eighteenth release (SR24) includes 7,906 food items and up to 146 food components [United States Department of Agriculture, 2011]. AGROVOC is a multi-lingual thesaurus developed by the FAO that has about 40,000 concepts and 3 types of relations (preferred term, related term and broader term) [FAO, 2011]. In 2002, FAO initiated action to enhance the quality of its information services related to fishery. Noy et al [2001] have developed an ontology to represent wine, food and appropriate combinations of wine with meals ontology. Graca et al. [2005] has proposed a specialized wine ontology coverings maceration, fermentation processes, grape maturity state, wine characteristics, and classification according to country and region where the wine was produced. Villarias
[2004] developed an ontology of culinary recipes to be used in a semantic querying system for the Web. A Food-Oriented Ontology-Driven System (FOODS) was presented by Snae et al. [2008] as a system for menu planning in a restaurant, clinic, hospital, or at home.

The library of the Central Food Technology Research Institute (CFTRI) also known as Food Science and Technology Information Service has an extensive collection of books, current periodicals, standards, patents and theses used by scientists and industries. The library brings out three publications namely Food Technology Abstracts (FTA), Food Patents (FP) and Food Digest (FD) enabling the research and development community to keep track of the technological advances in the area of FST worldwide. FST is a multidisciplinary area with several core and peripheral areas. Core subjects include food nomenclature, analysis, sensory evaluation, properties and processing. Peripheral subjects include chemical & physical sciences, engineering, technology and other related areas. Started in 1966, the monthly FTA focuses on FST research as well as review articles from various Indian and international journals. The other two services - FD and FP - are quarterly publications started during 1978 and 1980 respectively. While FD focuses on statistical information on production, processing, commercial intelligence and related news items, FP harvests national and international FST patent information. Together, the three constitute a knowledge base on R&D in FST.

Objective of the Study

The objective of this work is to represent the entire collection of the three information products, FTA, FP and FD in machine understandable form to create FOSTIS knowledge base (FSWKB). For this effective use of semantic web technologies is important. Once the representation of the entire knowledge base using an ontology is completed, the design and implementation of an inference engine with reasoning abilities will be initiated. The development of an intelligent software agent with options of taking user queries in a natural language form, transforming and processing the query using the inference engine to obtain answers to the queries is also planned.

Review of Methodologies

A clearly defined and well-structured methodology can reduce ontology development time and increase the possibility of success. Ontological engineering refers to the set of activities, the ontology life cycle, methods and methodologies for constructing ontologies, and tool suites and languages that support them (Gómez-Pérez et al., 2004). Mahesh (1996) proposed a situated methodology for ontology development and Grüninger & Fox (1995) proposed a methodology based on TOVE (TOronto Virtual Enterprise) project. The first activity in this methodology identifies the main scenarios that describe the purpose of the ontology with respect to the intended applications. Then, a set of competency questions are used to identify the scope of the ontology, thereby extracting the main concepts, properties, axioms of the underlying scope. After that, the elements of the ontology are expressed in first order logic. Uschold & King (1995) proposed a method built upon the experience assembled from developing an enterprise ontology. This method proposes the following activities: (i) identify the purpose of the ontology, (ii) build the ontology by capturing knowledge and identifying key concepts and properties in the domain, coding knowledge, and reusing other ontologies inside the current one, (iii) evaluate the ontology, and (iv) document the ontology.

In 1996, the methodology METHONTOLOGY (Gomez- Perez et al., 1996) for building ontologies from scratch or from reusing other ontologies was proposed (IEEE, 1996). It identifies ontology development process where the life cycle is based on evolving prototypes and particular techniques for carrying out each activity. The activities for building ontology are: project management activities including planning, control and
quality assurance; development-oriented activities including specification, conceptualization, formalization and implementation; support activities including acquisition, evaluation, integration, and documentation and configuration management. The current version of METHONTOLOGY (Lopez et al., 1999) can be included in this last generation of methodologies. In 2001, Noy and McGuinness proposed an iterative approach to ontology development which starts with a rough first pass at the ontology and incrementally refines it through interaction. It is also important to mention the study presented by Pinto and Martins [2004]. Corcho et al. (2005) suggested eleven specific tasks during the conceptualization stage of developing a conceptual ontology.

**Design of Ontology for FSWKB**

The FSWKB ontology is developed to enable researchers, scientists and other technical staff to semantically search the FST knowledge base. In order to conceptualize the ontology, domain analysis was performed to consider reusing existing ontologies and to acquire knowledge from the domain as the first step. The system is being developed in phases using an iterative approach.

**Enumeration of Glossary of Terms**

After analysis of the FST domain, a glossary of terms was built to include the entire vocabulary of FST (concepts, instances, and attributes). In order to standardise this glossary, some of the well-established ontology specifications, namely, KA2 initiative, ONTOWEB, Marcont, Dublin Core, HealthCyberMAP and FAO journal were studied. The glossary of terms is shown in Table 1.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resource</td>
<td>Base class for all bibliographical resources</td>
</tr>
<tr>
<td>Person</td>
<td>The class which represents the person</td>
</tr>
<tr>
<td>Language</td>
<td>A language of the intellectual content of the resource.</td>
</tr>
<tr>
<td>Subject</td>
<td>The topic of the content of the resource.</td>
</tr>
<tr>
<td>Topic</td>
<td>Topic of the resource</td>
</tr>
<tr>
<td>Organization</td>
<td>The organization that is involved in organizing an event or publishing given resource</td>
</tr>
<tr>
<td>Publisher</td>
<td>Describes publisher of a given resource</td>
</tr>
<tr>
<td>Access medium</td>
<td>Types of media available to access a resource</td>
</tr>
<tr>
<td>Article</td>
<td>Represents the concept of a scientific publication (usually in a journal).</td>
</tr>
<tr>
<td>Book</td>
<td>Represents the concept of a book</td>
</tr>
<tr>
<td>Book chapter</td>
<td>A chapter of a book</td>
</tr>
<tr>
<td>Collection</td>
<td>Represents collection of resources</td>
</tr>
<tr>
<td>In proceedings</td>
<td>One of the types of resources - an article in conference, workshop or proceedings</td>
</tr>
<tr>
<td>Manual</td>
<td>One of the types of resources - a manual or a use guide</td>
</tr>
<tr>
<td>Thesis</td>
<td>One of the types of resources - Master's thesis/PhD document</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>One of the types of resources which does not belong to any other category</td>
</tr>
<tr>
<td>News Item</td>
<td>Represents the latest news/happening</td>
</tr>
<tr>
<td>Patent</td>
<td>A document describing the exclusive right granted by a government to an inventor to manufacture, use, or sell an invention for a certain number of years</td>
</tr>
<tr>
<td>Lecture notes</td>
<td></td>
</tr>
<tr>
<td>Website</td>
<td>Represents a generic Internet-based publication medium</td>
</tr>
<tr>
<td>Title</td>
<td>The title of the work</td>
</tr>
<tr>
<td>Author/creator</td>
<td>Attaches information about author of the resource or the collection of resources</td>
</tr>
<tr>
<td>Name</td>
<td>Description</td>
</tr>
<tr>
<td>------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Contributor</td>
<td>An entity responsible for making contributions to the content of the resource.</td>
</tr>
<tr>
<td>Pages</td>
<td>Number of pages of the resource or chapter</td>
</tr>
<tr>
<td>Year/date</td>
<td>Date related to the entity</td>
</tr>
<tr>
<td>Format</td>
<td>The physical or digital manifestation of the resource.</td>
</tr>
<tr>
<td>Issue number</td>
<td>The number of a journal, magazine, technical report, or of a work in a series</td>
</tr>
<tr>
<td>Volume</td>
<td>The volume of a journal or multi-volume book</td>
</tr>
<tr>
<td>Address</td>
<td>Usually the address of the publisher or other type of institution</td>
</tr>
<tr>
<td>Abstract</td>
<td>Abstract of the resource</td>
</tr>
<tr>
<td>Type</td>
<td>The nature or genre of the content of the resource</td>
</tr>
<tr>
<td>URL</td>
<td>Property refers to a Digital Object Identifier assigned to a given resource</td>
</tr>
<tr>
<td>Keywords</td>
<td>Keyword related to the resource</td>
</tr>
<tr>
<td>Book title</td>
<td>Title of the book</td>
</tr>
<tr>
<td>Editor</td>
<td>Defines the editor of a given resource or collection of resources</td>
</tr>
<tr>
<td>ISSN/ISBN</td>
<td>Assigns an identifier in the form of ISSN/ISBN.</td>
</tr>
<tr>
<td>Edition</td>
<td>The edition of a book</td>
</tr>
<tr>
<td>Identifier (doi)</td>
<td>This property represents the unique identifier of the resource</td>
</tr>
<tr>
<td>Organization</td>
<td>Represent the concept of an organization</td>
</tr>
<tr>
<td>Pages From</td>
<td>The starting page of the given document (article) in the collection</td>
</tr>
<tr>
<td>Pages To</td>
<td>The ending page of the given document (article) in the collection</td>
</tr>
<tr>
<td>Source</td>
<td></td>
</tr>
<tr>
<td>Published</td>
<td></td>
</tr>
</tbody>
</table>

**Classes and their Hierarchy**

From the list in Table 1, we selected terms that describe objects having independent existence rather than terms that describe these objects. These terms are classes in the ontology and are anchors in the class hierarchy. Most of the remaining terms are properties of these classes. We organized the classes into a hierarchical taxonomy (Table 2).

**Table 2: Classes and their hierarchy in the FSWKB Ontology**

<table>
<thead>
<tr>
<th>RESOURCE</th>
<th>PERSON</th>
<th>ORGANIZATION</th>
<th>SUBJECT</th>
<th>THESAURUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article</td>
<td>o Staff</td>
<td>o Researcher</td>
<td>o Government (has to be evolved)</td>
<td>(has to be evolved)</td>
</tr>
<tr>
<td>Book</td>
<td>o Bookchapter</td>
<td>o Manual</td>
<td>o Collection</td>
<td>o In proceedings</td>
</tr>
<tr>
<td></td>
<td>o Thesis</td>
<td>o Master</td>
<td>o PhD</td>
<td>o Miscellaneous</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Properties of Classes-slots**

The classes alone will not provide enough information in order to answer the competency questions regarding the domain and scope of ontology. We have selected classes from the
list of terms that are created in Table 2 for describing the internal structure of concepts. For each property in the list, we have determined which class it describes. These properties become slots attached to classes. Table 3 shows the properties of classes-slots

**Table 3: Properties of classes-slots in the FSWKB Ontology**

<table>
<thead>
<tr>
<th>Sl. No</th>
<th>Class name</th>
<th>Class attributes</th>
<th>Instance Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Resource</td>
<td>creator, editor, subject, thesaurus</td>
<td>Title, Year, abstract, identifier (URL), Note, contents, contributor, format, publisher, language</td>
</tr>
<tr>
<td>1.1</td>
<td>Article</td>
<td></td>
<td>Journal-title, Pages To, Pages From, volume, issue number, issn, doi</td>
</tr>
<tr>
<td>1.2</td>
<td>Book</td>
<td></td>
<td>Book-title, edition, issn</td>
</tr>
<tr>
<td>1.2.1</td>
<td>Book chapter</td>
<td></td>
<td>Doi, Pages From, Pages To</td>
</tr>
<tr>
<td>1.2.2</td>
<td>Manual</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>1.3</td>
<td>Collection</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>1.4</td>
<td>in proceedings</td>
<td></td>
<td>organizer Conference name, date, location</td>
</tr>
<tr>
<td>1.5</td>
<td>Thesis</td>
<td>Contributor, university</td>
<td>Degree, type</td>
</tr>
<tr>
<td>1.5.1</td>
<td>Master Thesis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.5.2</td>
<td>Phd. Thesis</td>
<td></td>
<td>Faculty</td>
</tr>
<tr>
<td>1.6</td>
<td>News Item</td>
<td></td>
<td>Newspaper name, date, pages</td>
</tr>
<tr>
<td>1.7</td>
<td>Patent</td>
<td>Inventor, assignee</td>
<td>Application number, Patent number, country</td>
</tr>
<tr>
<td>1.9</td>
<td>Lecture notes</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>1.10</td>
<td>Website</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>1.11</td>
<td>Miscellaneous</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>Person</td>
<td>Organization, Subject area,</td>
<td>first-name, last-name, address, e-mail, department, joined-year, qualification, Gender</td>
</tr>
<tr>
<td>2.1</td>
<td>Researcher</td>
<td></td>
<td>Has Role no,</td>
</tr>
<tr>
<td>2.2</td>
<td>Staff*</td>
<td>competency</td>
<td>Has-employee no, Has-role, has- URL</td>
</tr>
<tr>
<td>2.3</td>
<td>Student*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Organization</td>
<td>affiliated organization, sub organization</td>
<td>Mission, name, has Url, address, e-mail, has employees, orgn., type</td>
</tr>
<tr>
<td>4</td>
<td>Thesaurus</td>
<td></td>
<td>Keyword, subtopic, super topic, preferred term, descriptor for, entry term for About subject,</td>
</tr>
<tr>
<td>5</td>
<td>Subject Scheme</td>
<td></td>
<td>Keyword, Key phrases Classification codes</td>
</tr>
</tbody>
</table>

**Create Instances**

In this step individual instances of classes in the hierarchy were created. Table 4-8 shows an instance created for individual classes in the ontology.

**Table 4: Instance Government for the class Organisation**

<table>
<thead>
<tr>
<th>Instance name</th>
<th>Class attribute</th>
<th>Instance attribute</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSIR</td>
<td>Affiliated organization=&quot;null&quot;</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sub-organization=&quot;CFTRI&quot; (Instance of organisation)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CFTRI</td>
<td>Affiliated organization=&quot;csir&quot;</td>
<td>Mission=&quot;Sustain leadership in long-term strategic Res.tech.dev&quot; affiliation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sub-organization=&quot;null&quot;</td>
<td>Name=&quot;CFTRI&quot;</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Address=&quot;Mysore,Karnataka&quot;</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>e-mail=&quot;<a href="mailto:fostis@cftri.res.in">fostis@cftri.res.in</a>&quot;</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>orgn_type=&quot;Central&quot;</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>URL=www.cftri.com</td>
<td></td>
</tr>
</tbody>
</table>
### Table 5: Instance staff for the class Person

<table>
<thead>
<tr>
<th>Instance</th>
<th>Class attribute</th>
<th>Instance attribute</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kalpana Patel (staff instance)</td>
<td>Organisation=&quot;CFTRI&quot; Subject=&quot;Food-1&quot;</td>
<td>First-name=&quot;Kalpana&quot; Last-name=&quot;Platel&quot; e-mail=<a href="mailto:biochem@cftri.res.in">biochem@cftri.res.in</a> Department=&quot;B&amp;N&quot; Joined-year=&quot;1980&quot; Qualification=&quot;M.Sc.&quot; Emp.No=&quot;1934&quot; Role=&quot;Senior Scientist&quot;</td>
<td>Organisation subject</td>
</tr>
<tr>
<td>Smita Gautam (Researcher instance)</td>
<td>Organisation=&quot;CFTRI&quot; Subject=&quot;Food-1&quot;</td>
<td>First-name=&quot;Smita&quot; Last-name=&quot;Gautam&quot; e-mail=<a href="mailto:biochem@cftri.res.in">biochem@cftri.res.in</a> Department=&quot;B&amp;N&quot; Joined-year=&quot;1980&quot; Qualification=&quot;M.Sc.&quot; RoleNo=&quot;RF0123&quot;</td>
<td>Organisation subject</td>
</tr>
</tbody>
</table>

### Table 6: Instance for class Thesaurus

<table>
<thead>
<tr>
<th>Instance</th>
<th>Class attribute</th>
<th>Instance attribute</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term-1</td>
<td></td>
<td>Keyword=&quot;Food additives&quot; Supertopic=&quot;additives&quot; Subtopic=&quot;sweeteners&quot;</td>
<td></td>
</tr>
</tbody>
</table>

### Table 7: Instance for class Subject scheme

<table>
<thead>
<tr>
<th>Instance</th>
<th>Class attribute</th>
<th>Instance attribute</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food-1</td>
<td></td>
<td>Keyword=&quot;Food Technology&quot; Classification codes=&quot;F8,3&quot;</td>
<td></td>
</tr>
</tbody>
</table>

### Table 8: Instance of Article (subclass of resource)

<table>
<thead>
<tr>
<th>Instance name</th>
<th>Class attribute</th>
<th>Instance attribute</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kalpana Patel Et al</td>
<td>Subject=&quot;Food-1&quot; Creator=&quot;Kalpana Patel, Smita Gautam&quot; Thesaurus=&quot;Term-1&quot;</td>
<td>Title=&quot;Influence of combinations of promoter and inhibitor on the bioaccessibility of iron and zinc from food grains&quot; Year=&quot;2011&quot; Abstract=&quot;Inherent phytic acid and tannins interfere with………&quot; Identifier(URL)=www.informahealthcare.com Publisher=&quot;Informa Healthcare&quot; Language=&quot;English&quot; Format=&quot;pdf&quot; Journal-title=&quot;Intl JI of Food Sci. &amp; Nutr&quot; Volume=&quot;62&quot; Issnumber=&quot;8&quot; PagesFrom=&quot;826&quot; PagesTo=&quot;834&quot; issn=&quot;1465-3478&quot; Doi=&quot;10.3109/09637486.2011.584861&quot;</td>
<td>subject creator thesaurus</td>
</tr>
</tbody>
</table>
Conclusion

In this paper, we have described the general framework for the design of the ontology for the FSWKB. In order to arrive at this specification, we have analysed various existing ontology design methodologies. There is no one correct way or methodology for developing ontologies. There are several viable alternatives to model a domain. The best solution almost always depends on the application and domain specification requirements. Ontologies are generally developed using an iterative process to arrive at a good design. The first prototype of the FSWKB ontology is presented in this paper and needs formal representation using an ontology language OWL. This will be undertaken in our future work.

The ontology is expected to evolve by including the formalization of each term and the constraints on its properties. Terms will be represented through classes, relations, functions, and instances. To implement our FSWNB, we have chosen OWL-DL as an ontology development language and Protégé-OWL as an implementation tool.

Acknowledgement

The authors thank Dr. G. Venkateswara Rao, Acting Director, CFTRI, for his support and encouragement and Mr. Sanjai Lal, Technical Assistant, FOSTIS, CFTRI for his useful suggestions.

References


MarcOnt Bibliographic Ontology, MarcOnt Bibliographic Ontology 2.1 Specification, http://semdl.info/books/2/appendices/G#SECTION07720061000000000000


Sangeeta Deokattey - Bhabha Atomic Research Centre, Mumbai
D.K. Dixit - Bhabha Atomic Research Centre, Mumbai
K. Bhanumurthy - Bhabha Atomic Research Centre, Mumbai

Co-word and Facet Analysis as Tools for Conceptualization in Ontologies:
A Preliminary Study of a Micro-domain

Abstract
Conceptualization is at the core of developing domain ontologies. This paper reports a study for developing an ontology for a micro-domain - Test Blanket Module (TBM), an integral part of thermonuclear or fusion reactors. Sample data downloaded from yielded 1115 unique DEI (indexer-assigned) descriptors assigned to 548 records on TBM. The frequencies of occurrence of all the unique descriptors, the corresponding co-word DEI descriptors (AN numbers) were identified On the basis of their research linkages the descriptors were grouped into four quadrants. It was found, that the descriptors in the 2nd and 3rd quadrants were at the core of the selected subject. A total of 31 core descriptors from these were selected for conceptualization and for each the co-occurring descriptors and their frequencies of co-occurrence with the selected descriptor were noted. Only descriptor pairs that co-occurred 10 times or higher were considered. Comparison of Co-Word Word Blocks (CWWBs) and word blocks (INISWB) from the INIS thesaurus showed differences. Co-words were used to semantically enrich descriptors transforming them into more comprehensive concepts; these were used as building blocks for conceptualization and for domain ontology. This method could be replicated to generate semantic networks (which could form an Ontological layer on any subject of study) and also in query expansion during search and retrieval in interdisciplinary subject domains.

Introduction
Domain Ontologies are being developed to improve retrieval in the digital environment. According to Chandrasekaran (1999) an ontology is a representative vocabulary of a domain. More precisely, it is not the vocabulary as such that qualifies as an ontology, but the conceptualization that the terms in the vocabulary, are intended to capture. An early experiments to convert a controlled vocabulary into an Ontology was reported by Qin and Paling (2001). They used ERIC descriptors to develop an ontology for Education. According to them, the major difference between the two models is in the value addition through deeper semantics, in describing digital objects, both conceptually and relationally. A study by Poli (2002) highlights ontological sub-theories and the use of domain analysis for developing an ontology. Prieto-Diaz (2003) also used domain analysis and a faceted approach to build Ontologies using the ‘DARE’ software. Most of the current ontology projects use readily-available tools for developing new ontologies. Using the principles of latent semantic analysis, Zheng (2009) et al. reported the automatic generation of ontologies through conceptualization. Methontology (a well structured methodology used to build ontologies from scratch) and Facet analysis approaches were reported by Garcia-Torres et al. (2008) to develop an ontology. A study on automatic mining of domain concepts from documents was done by Cho-Wei Shih et al. (2011) using the crystallization method. The OntoPlus methodology was proposed by Novalija et al. (2011) for ontology extension using automatic text mining and co-occurrence information.
Materials and Method

A micro-domain, viz., Test Blanket Module (TBM) was selected for the present study. TBMs are an integral part of thermonuclear or fusion reactors. Sample data of 548 bibliographic records for the study (2001-2010) was downloaded from the INIS database using the following queries:

**Search History**

- #4 #1 or #3 (548 records)
- #3 ((ITER-TOKAMAK) in DE) and (PY=2001-2010) and TBM and (PY=2001-2010) (387 records)
- #2 ((ITER-TOKAMAK) in DE) and (PY=2001-2010) (6591 records)
- #1 TBM and (PY=2001-2010) (548 records).

Apart from the usual fields in any bibliographic record, each bibliographic record contained two kinds of descriptors from the INIS thesaurus: DEI (Indexer assigned descriptors) and DEC (Computer assigned descriptors) and a unique Accession Number (AN). A total of 1115 unique DEI descriptors and 606 unique DEC descriptors were found assigned to the 548 records. The descriptors assigned by human indexers (DEI) are more specific and representative of the semantic content and therefore, DEI descriptors were selected for the study. The entire text file was parsed and filtered to separate the DEI field and then each DEI descriptor and its corresponding AN. The frequency of occurrence of each unique DEI descriptor and the corresponding co-occurring DEI descriptors (AN numbers) were identified and a single file of each DEI descriptor, its frequency of occurrence in the entire text file and other co-occurring DEI descriptors (AN numbers) associated with it, was generated. Java and MySQL were used for this process. When these descriptors were grouped into four quadrants on the basis of the centrality axis of their research linkages, it was found, that the top ranking descriptors were too general in scope and were thus not included for co-word and facet analysis. The descriptors in the 2nd and 3rd quadrants were found to be at the core of the subject. Descriptors with a low level of research linkages were in the 4th quadrant. Descriptors with a threshold of less than 20 were excluded from the study. Thus a total of 31 core descriptors from the 2nd and 3rd quadrants were identified for the final analysis (Table 1).

<table>
<thead>
<tr>
<th>Table1: Break-up of the DEI descriptors for sample data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of unique DEI descriptors found for the 548 records on TBM</td>
</tr>
<tr>
<td>Total number of highest ranking descriptors (1st quadrant)</td>
</tr>
<tr>
<td>Total number of descriptors with a threshold level of &lt;20</td>
</tr>
<tr>
<td>Total number of descriptors in the 2nd and 3rd quadrants</td>
</tr>
<tr>
<td>Total number of descriptors in the 4th quadrant</td>
</tr>
</tbody>
</table>

Each of the 31 selected descriptors was arranged alphabetically (with their individual frequencies of occurrence) along with the other descriptors and their frequencies of co-occurrence. Here too, a threshold level of 10 and above was maintained. A descriptor pair which co-occurred less than 10 times was not considered for the analysis. A sample of a descriptor and its co-word pairs is given in Fig. 1.

Only two instances of co-occurrence were selected for further facet analysis, viz., *thermonuclear reactor materials and beryllium (14)* and *thermonuclear reactor materials and breeding blankets (68)*. Similar CWWBs were generated for each of the 31 descriptors. Similarly, a thesaurus word block (INISWB) for all the selected descriptors was also downloaded from the INIS online thesaurus. A sample of one of the selected descriptors with both the word blocks is shown in Fig 2.
### Fig. 1

**THERMONUCLEAR-REACTOR-MATERIALS**

<table>
<thead>
<tr>
<th>Material</th>
<th>Frequency of Occurrence: 90</th>
</tr>
</thead>
<tbody>
<tr>
<td>accelerator-facilities</td>
<td>An:40-015618</td>
</tr>
<tr>
<td>accelerators</td>
<td>An:40-015620</td>
</tr>
<tr>
<td>atomic-displacements</td>
<td>An:34-013155, 37-110165, 40-10936,41-006922,41-6927,42-041073</td>
</tr>
<tr>
<td>beryllium-allys</td>
<td>An:39-015460, 39-015644,42-041073</td>
</tr>
<tr>
<td>bonding</td>
<td>An:36-008322,36-080647, 38-005626</td>
</tr>
<tr>
<td>boron</td>
<td>An:37-110165</td>
</tr>
<tr>
<td>boundary-conditions</td>
<td>An:39-010081, 39-015534</td>
</tr>
</tbody>
</table>

### Fig. 2

**FERRITIC STEELS**

<table>
<thead>
<tr>
<th>Material</th>
<th>Frequency of Occurrence: 84</th>
</tr>
</thead>
<tbody>
<tr>
<td>INISWB</td>
<td></td>
</tr>
<tr>
<td>BT1 steels</td>
<td></td>
</tr>
<tr>
<td>NT1 steel-cr12moniv</td>
<td></td>
</tr>
<tr>
<td>NT1 steel-cr13al</td>
<td></td>
</tr>
<tr>
<td>NT2 stainless-steel-405</td>
<td></td>
</tr>
<tr>
<td>NT1 steel-cr16</td>
<td></td>
</tr>
<tr>
<td>NT2 stainless-steel-430</td>
<td></td>
</tr>
<tr>
<td>NT1 steel-cr25</td>
<td></td>
</tr>
<tr>
<td>NT2 stainless-steel-446</td>
<td></td>
</tr>
<tr>
<td>NT1 steel-cr9mo</td>
<td></td>
</tr>
<tr>
<td>NT1 steel-cr9monb</td>
<td></td>
</tr>
<tr>
<td>RT corrosion resistant alloys</td>
<td></td>
</tr>
<tr>
<td>RT ferrite</td>
<td></td>
</tr>
<tr>
<td>CWWB</td>
<td></td>
</tr>
<tr>
<td>Tritium 29*</td>
<td></td>
</tr>
<tr>
<td>Thermonuclear-reactor-materials 26</td>
<td></td>
</tr>
</tbody>
</table>

(*The numbers in the CWWB column are the frequencies of co-occurrence of the respective descriptors with the main descriptor*)
Analysis and Results

The sample data of 548 records on TBM downloaded from the INIS database contained 1115 unique indexer-assigned descriptors. The number of unique Computer assigned descriptors (DEC) for the 548 records was 606. The five highest ranking DEI & DEC descriptors are shown in Table 2.

Table 2: Top ranking DEI and DEC descriptors on TBM

<table>
<thead>
<tr>
<th>DEI Descriptors</th>
<th>Frequency of Occurrence</th>
<th>DEC Descriptors</th>
<th>Frequency of Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>iter-tokamak</td>
<td>387</td>
<td>thermonuclear-devices</td>
<td>391</td>
</tr>
<tr>
<td>breeding-blankets</td>
<td>301</td>
<td>thermonuclear-reactors</td>
<td>388</td>
</tr>
<tr>
<td>helium</td>
<td>203</td>
<td>tokamak-devices</td>
<td>387</td>
</tr>
<tr>
<td>design</td>
<td>173</td>
<td>tokamak-type-reactors</td>
<td>387</td>
</tr>
<tr>
<td>tritium</td>
<td>169</td>
<td>closed-plasma-devices</td>
<td>331</td>
</tr>
<tr>
<td></td>
<td></td>
<td>reactor-components</td>
<td>304</td>
</tr>
</tbody>
</table>

Seventeen unique DEI descriptors from the 2nd and 3rd quadrants were not selected for co-word and facet analysis as they were very general in scope and not related to the core R&D areas on TBM. During co-word analysis it was observed that there was not a single instance of repetition of the descriptors in the INIS Word block and the Co-Word Word block in the 31 selected DEI descriptors, thus providing ample scope for further facet analysis. Only three kinds of relationships were seen among the descriptors in the INIS Word block: Equivalent, Hierarchical and Associative and all falling under the Personality facet. An example from the descriptor ferritic-steels is shown in Table 4. In the Co-word word block of the descriptor, all the relationships were associative; higher the frequency of co-occurrence, stronger is the affinity between the two descriptors. For instance, ferritic-steels and breeding blankets co-occurred 40 times, stressing the importance of ferritic-steels as structural materials used in the breeding blankets of TBM. Table 4 provides the details.

Table 3: Types of relationships seen in a descriptor from the INIS Word block

<table>
<thead>
<tr>
<th>FERRITIC-STEELS</th>
<th>Freq.of Occur. 84</th>
</tr>
</thead>
<tbody>
<tr>
<td>NT1 steel-cr12moniv</td>
<td>INIS</td>
</tr>
<tr>
<td>NT1 steel-cr13al</td>
<td>INIS</td>
</tr>
<tr>
<td>NT2 stainless steel-405</td>
<td>INIS</td>
</tr>
<tr>
<td>NT1 steel-cr16</td>
<td>INIS</td>
</tr>
<tr>
<td>NT2 stainless steel-430</td>
<td>INIS</td>
</tr>
<tr>
<td>NT1 steel-cr25</td>
<td>INIS</td>
</tr>
<tr>
<td>NT2 stainless steel-446</td>
<td>INIS</td>
</tr>
<tr>
<td>NT1 steel-cr9mo</td>
<td>INIS</td>
</tr>
<tr>
<td>NT1 steel-cr9monhv</td>
<td>INIS</td>
</tr>
<tr>
<td>RT corros-resistant alloys</td>
<td>INIS</td>
</tr>
<tr>
<td>RT ferrite</td>
<td>INIS</td>
</tr>
</tbody>
</table>

Each of the Associative relationships was analyzed using facet analysis and the type of semantic relationship between the two descriptors was specified. The Generalized Facet Structure (GFS) framework based on the General Theory of Classification of S.R. Ranganathan further elaborated by Neelameghan (1975) was used, with some modifications, as a guideline for conceptualization.
### Table 4: Types of relationships seen in a descriptor from the Co-word Word block

<table>
<thead>
<tr>
<th>FERRITIC STEELS</th>
<th>Fr Oc.</th>
<th>Occur</th>
<th>Source</th>
<th>Type of relationship</th>
<th>Facet</th>
</tr>
</thead>
<tbody>
<tr>
<td>tritium</td>
<td>29</td>
<td>CWWB</td>
<td>Associative</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>thermonuclear-reactor-materials</td>
<td>26</td>
<td>CWWB</td>
<td>Associative</td>
<td>P/M</td>
<td></td>
</tr>
<tr>
<td>atomic-displacements</td>
<td>10</td>
<td>CWWB</td>
<td>Associative</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>beryllium</td>
<td>11</td>
<td>CWWB</td>
<td>Associative</td>
<td>P/M</td>
<td></td>
</tr>
<tr>
<td>breeding-blankets</td>
<td>40</td>
<td>CWWB</td>
<td>Associative</td>
<td>P/M</td>
<td></td>
</tr>
<tr>
<td>coolants</td>
<td>20</td>
<td>CWWB</td>
<td>Associative</td>
<td>P/M</td>
<td></td>
</tr>
<tr>
<td>creep</td>
<td>10</td>
<td>CWWB</td>
<td>Associative</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>design</td>
<td>37</td>
<td>CWWB</td>
<td>Associative</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>fatigue</td>
<td>10</td>
<td>CWWB</td>
<td>Associative</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>first-wall</td>
<td>23</td>
<td>CWWB</td>
<td>Associative</td>
<td>P/M</td>
<td></td>
</tr>
<tr>
<td>helium</td>
<td>27</td>
<td>CWWB</td>
<td>Associative</td>
<td>P/M</td>
<td></td>
</tr>
<tr>
<td>irradiation</td>
<td>13</td>
<td>CWWB</td>
<td>Associative</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>iter-tokamak</td>
<td>69</td>
<td>CWWB</td>
<td>Associative</td>
<td>P/M</td>
<td></td>
</tr>
<tr>
<td>liquid-metals</td>
<td>11</td>
<td>CWWB</td>
<td>Associative</td>
<td>P/M</td>
<td></td>
</tr>
<tr>
<td>lithium</td>
<td>18</td>
<td>CWWB</td>
<td>Associative</td>
<td>P/M</td>
<td></td>
</tr>
<tr>
<td>martensitic-steels</td>
<td>40</td>
<td>CWWB</td>
<td>Associative</td>
<td>P/M</td>
<td></td>
</tr>
<tr>
<td>microstructure</td>
<td>11</td>
<td>CWWB</td>
<td>Associative</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>neutrons</td>
<td>17</td>
<td>CWWB</td>
<td>Associative</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>performance</td>
<td>10</td>
<td>CWWB</td>
<td>Associative</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>breeding-blankets</td>
<td>12</td>
<td>CWWB</td>
<td>Associative</td>
<td>P/M</td>
<td></td>
</tr>
</tbody>
</table>

Concept maps were generated for each of the 31 selected descriptors along with their word blocks using the CMap Tools Knowledge Modeling Kit. A sample concept map on Ferritic steels created using the CMap tools is shown in Fig. 3. The linkages which are displayed through thick lines and the corresponding descriptors in italics, are from the co-word word block. Thus each of the 31 descriptors was transformed into a concept. They were then linked with each other to generate an interlinked concept map on TBM. Under the guidance of subject experts, major concepts on TBM were identified both from the DEI descriptors and others separately and this interlinked concept map was merged with them to form a domain ontology layer on TBM.

**Fig. 3: A concept map of one of the selected descriptors**

![Concept Map of Selected Descriptor](image-url)
Conclusions
Developing domain ontologies in interdisciplinary and multidisciplinary areas of R&D is a challenge to Information Scientists. First there are no standard vocabulary control tools available for such narrow micro domains. Second, creating new domain ontologies is equally challenging and time-consuming. The method employed in this preliminary study is based on reuse of existing vocabulary control tool, a thesaurus. It also made use of standard, KO&R methods of co-word and facet analysis for conceptualization. It showed how descriptors of a thesaurus can be converted into concepts, thus making conceptualization easy. Apart from the controlled vocabulary descriptors of the INIS thesaurus used in this study, free-text terms from author abstracts can also be incorporated to make the semantic network more dynamic and comprehensive. This is necessary as thesaurus updates are not very frequent and author abstracts can provide valuable information in a rapidly developing subject.

References
A Theoretical Framework for Operationalizing Basic Level Categories in Knowledge Organization Research

Abstract

Research on categories indicates that superordinate categories lack informativeness because they are represented by only a few attributes while subordinate categories lack cognitive economy because they are represented by too many attributes (e.g., Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Basic level categories balance informativeness and cognitive economy: They represent the most attributes common to category members and the fewest attributes shared across categories. Green (2006) has suggested that the universality of basic level categories can be used for building crosswalks between classificatory systems. However, studies of basic level categories in KO systems have assumed that the notion of a basic level category is understood and have failed to operationize the notion of "basic level category" before applying it in the analysis of user-generated vocabularies. Heidegger's (1953/1996) notion of handiness (i.e., zuhandenheit, or being "at hand") can provide a framework for understanding the unstable and relational nature of basic level categories and for operationalizing basic level categories in KO research.

A substantial body of research indicates that basic level categories are the first categories formed during perception of the environment, the first learned by children, and those most used in language (e.g., Mervis & Crisafi, 1982; Tversky & Hemenway, 1983). In his seminal article, Roger Brown (1958), who is often identified as the first researcher to have theorized the existence of a basic level of categorization (e.g., Lakoff, 1986; Murphy 2002), described how parents teach their children to name and categorize objects. He argued that the "name of a thing, the one that tells what it 'really' is, is the name that constitutes the referent as it needs to be constituted for most purposes" (p. 17). Brown points out that, while some names given to an object are "recategorizations" that represent "acts of imagination" (p. 17), the names of things as they really are serve as true representations of the referent. Brown's contention that category names of things represent things "as they are" -- as the thing "needs to be constituted for most purposes" -- foreshadowed Eleanor Rosch's theory of basic level categories.

Drawing inspiration from Brown's (1958) observations and from Wittgenstein's (1953/1963) discussions of language use, Rosch and her colleagues (Rosch & Mervis, 1975; Rosch et al., 1976) developed the theory of basic level categories. Rosch (1978) argued that, in the perceived world, information-rich bundles of perceptual and functional attributes form natural discontinuities and that basic cuts in categorization are made at these discontinuities (p. 31). Working from these assumptions, Rosch et al. (1976; Rosch, 1978) defined basic level categories as most inclusive because they convey "the most information, possess the highest cue validity, and are, thus, the most differentiated from one another" (Rosch et al., 1976, p. 383). According to Rosch and Mervis (1975):

Basic objects (for example, chair, car) are the most inclusive level of abstraction at which categories can mirror the correlational structure ... of the environment and the most inclusive level at which there can be many attributes common to all or most members of the categories. (p. 586)

Thus, while experimental studies have suggested that superordinate categories lack informativeness because they are represented by only a few attributes and that subordinate categories lack cognitive economy because they are represented by too many attributes (Rosch & Mervis, 1975), basic level categories provide a balance between informativeness and cognitive economy because they represent the most inclusive level of abstraction: the level at which the greatest number of attributes are common to category members (e.g.,
dog) and the fewest number of attributes are shared across categories (e.g., dog and cat). In the years since Rosch introduced the notion of basic level categories, her theory has been applied in a wide variety of cognitive and linguistic studies, including word use and free naming (e.g., Tanaka & Taylor, 1991; Markman & Wisniewsky, 1997; Blewitt, 1983); American Sign Language for the deaf (e.g., Newport & Bellugi, 1978); categorization of images (e.g., Murphy & Smith, 1982); and environmental scenes (e.g., Tversky & Hemenway, 1983; 1984).

Studies of basic level categories have indicated that entry level categorization (i.e., the level of abstraction that is perceived or identified first during the individual's interactions with the environment) may be modulated by two factors: the individual's domain-specific knowledge and the typicality of an exemplar for a corresponding category. Thus Jolicoeur et al. (1984) claim that the "level at which objects are identified first depends on typicality" (p. 271), while Rips et al. (2006) speculate that factors responsible for an individual's preference for a basic level term "may include not only those associated with its basic-level category but also the larger set of background causes that govern the individual's environment" (p. 9). It has frequently been argued that both the influence of an individual's experience and the typicality of the object within a given environment may actually shift the entry level of categorization to the subordinate level (e.g., Rosch et al., 1976; Jolicoeur et al., 1984; Tanaka & Taylor, 1991; Belke et al., 2010). As Brown (1958) observed:

"It seems likely that things are first named so as to categorize them in a maximally useful way. For most purposes Referent A is a spoon rather than a piece of silverware, and Referent B a dime rather than a metal object. The same referent may have its most useful categorization on one level (Prince) for one group (the family) and on another level (dog) for another group (strangers). The categorization that is most useful for very young children (money) may change as they grow older (dime and nickel)." (p. 20; emphasis in original)

Within the framework of the experientialist view, conceptual structure is meaningful because it is embodied, that is, it arises from, and is tied to, our perceptual bodily experiences" (Lakoff, 1987, p. 267). For Lakoff, embodiment is coupled with "collective biological capacities and ... physical and social experiences [of individuals] as beings functioning in ... [their] environment" (p. 267). The experientialist view assumes the existence of an external structure in "our bodily experience" (p. 267), which arises from the structured "preconceptual experiences" (p. 267) of individuals. Lakoff (see also Lakoff & Johnson, 1980) suggests that there are at least two kinds of structure that are implicated in preconceptual experiences: basic level structure and kinesthetic image-schematic structure. Basic level structures are comprised of categories that represent the convergence of gestalt perception, the capacity for bodily movement, and the development of rich mental images, while kinesthetic image-schematic structures are the external structures that constantly appear in everyday bodily experiences -- structures such as paths, containers, or front-back and part-whole relationships (p. 267). Lakoff also theorizes that, through the use of metaphors and the projection of basic level categories to superordinate and subordinate categories, individuals extend these basic level and image-schematic structures to create abstract conceptual structures. However, the experientialist approach does not address the relational and unstable nature of preconceptual structures such as basic level categories: For example, the perception and understanding of categories and category boundaries can differ substantially across individuals, situations, and cultures and even change dramatically over time, and the experientialist view cannot account for such variability and instability.

LIS research has tended to confine research on basic level categories to two areas of investigation: How the basic level of categories is represented in knowledge organization systems (e.g., Iyer, 1995; Fernandez & Eastman, 1990; Green, 2006); and how basic level terms are used to represent the conceptual content of resources such as images (e.g.,
Jorgensen, 2003; Rorissa & Iyer, 2008; Rorissa, 2008; Yoon, 2009; Schmidt & Stock, 2009). Green, Bean and Hudon (2002) investigated the occurrence of basic level categories in four different systems of knowledge organization: a bilingual thesaurus (Canadian Literacy Thesaurus), a biomedical vocabulary (Unified Medical Language System), and two ontologies (ThoughtTreasure and WordNet). Analysis of a random sample of terms taken from these resources, which included the semantic scope of a term as well as its placement in the original hierarchy, demonstrated that equivalence across systems occurred significantly at the basic level rather than at subordinate or superordinate levels. Interestingly, basic level terms were observed most often in the Canadian Literacy Thesaurus and least often in the Unified Medical Language System. Analysis using goodness-of-fit tests confirmed that, in these four systems, basic level terms were more likely to have equivalent terms across systems than were subordinate or superordinate terms.

The most extensive study of basic level categories reported in the LIS literature was conducted by Green (2006). She identified a set of basic level terms by analyzing the structure of WordNet, a lexical database of English words in which 117,000 groups of synonyms (or synsets) are linked by hierarchical (i.e., superordinate-subordinate) or part-whole relationships. Based on her analysis, Green (2006, p. 8) developed six general criteria for the identification of basic level categories:

- Length and structure of lexical units (i.e., whether the lexical unit is a simple word, a compound word, or a phrase);
- Level of occurrence within the WordNet tree structure;
- Number of links to parts;
- Frequency of usage;
- Total number of links to other categories; and
- Number of immediate children (i.e., subordinate categories one level down) and overall number of children (all subordinates).

Applying these criteria to an analysis of every leaf node in WordNet's network of 59,692 noun synsets, Green identified 7,168 basic level categories.

To investigate the hypothesis that category terms at the basic level were more universal and therefore more likely to co-occur across classificatory systems than either subordinate or superordinate category terms, Green investigated the occurrence of basic level categories in a sample of 28 thesauri in ten subject domains (i.e., agriculture, education, engineering, environment, graphic materials, health, information science, legislation, political science, population science, and water sciences). She randomly selected ten category terms from each thesaurus and expanded these terms using the relational structure of the thesaurus from which each term was drawn. For each category term in these hierarchical expansions, the closest corresponding category term was identified in each of the other thesauri for that subject domain to produce category pairs consisting of one category term from the hierarchical expansion in one thesaurus and the closest corresponding term in a second thesaurus. Green then analyzed all category pairs to determine if the two category terms were equivalent and applied the criteria for basic level categories generated in the analysis of WordNet to identify basic level category terms.

Green's findings appear to support the hypothesis that basic level terms are universal. She suggests that the practical outcome of her study is the direction it provides for building crosswalks between classificatory systems: Given that basic level categories are significantly more likely to have exact equivalents across systems than categories not at the

---

1. [http://wordnet.princeton.edu](http://wordnet.princeton.edu)
basic level, developing mappings across classificatory systems should emphasize basic level categories to generate the "cleanest mappings" (2006, p. 12). However, even though the apparent universality of basic level categories in systems of knowledge organization and representation would seem to indicate their universality in indexing as well, basic level categories may not be appropriate for supporting collocation precisely because they have such a high probability of occurrence.

Applying the experientialist approach, Tennis (2005) proposed a user-centered classificatory structure that could potentially improve the static nature of traditional classification systems through incorporation of user-generated terms that are linked to the concepts defined in classification system. However, Tennis's proposal inherits the problems of the experientialist approach, including, in particular, the unstable and relational nature of an individual's preconceptual structure. Moreover, the instability and variability of individual preconceptual structures may also affect the process of classification and the application of principles of classification that govern the structure and relationship of classes (Jacob, 2004).

Although the universality of basic level terms is generally accepted by researchers in LIS, many of whom have suggested the positive outcomes of considering basic level terms in information representation and organization (Iyer, 1995; Bates, 1998; Hoenkamp, Stegeman & Schomaker, 1999), few studies have actually tested the utility of basic level terms in information organization and retrieval. Furthermore, studies that have investigated basic level terms generally have not provided an account of how the notion of basic level categories was actually operationalized (e.g., Golder & Huberman; Munk & Mork, 2007). Several studies (e.g., Rorissa & Iyer, 2008; Rorissa, 2008; Schmidt & Stock, 2009) have adapted the coding scheme that Green devised in her analysis of WordNet (Green et al., 2002; Green, 2006). However, even though WordNet has become the "de facto standard because it is freely available for research" (Ide & Wilks, 2007, p. 52; emphasis in original), the database itself was constructed based on paradigmatic relationships between nouns and verbs, which can lead to difficulties when attempting to determine the exact semantic level of a particular term. Thus, for example, because WordNet does not include adjectives, Schmidt and Stock (2009) were forced to identify the noun form to which an adjective was related in order to identify hierarchical relationships (e.g., to identify happy with the hierarchical position of happiness).

Problems associated with Green's (2006) operationalization of basic level terms based on the hierarchical structure of synsets in WordNet are not limited to the determination of a term's exact semantic level. More importantly, this approach masks the importance of the context in which a category term occurs and thus fails to take into account the relational nature -- the "handiness" or Zuhandenheit (Heidegger, 1953/1996, p. 65) -- of basic level terms.

Because the relational nature of basic level categories is either masked or ignored when the research focus is on the category as a term, Heidegger's notion of handiness offers critical insights for more effective operationalization of basic level categories. Heidegger defines "handiness" -- the notion of being "at hand" or "ready-to-hand" -- as the "ontological categorical definition of beings as they are, in themselves" (p. 67; emphasis in original). Handiness is the pre-reflective understanding of a thing as useful, or "ready-to-hand", on the basis of what is objectively present" (p. 67) -- on the basis of a thing as it is and how it is or, as Brown (1958) observed, "how it needs to be constituted for most purposes" (p. 17). In Heidegger's account, handiness is the subject of assertion which ultimately plays some role in practical inference" (Rorty, 1991, p. 32) and thus reflects and contributes to the individual's understanding of the world as it is. Thus handiness is an
assertion that is grounded in inference based on the realization of things as they are. In sharp contrast to the Platonic account, which limits assertion to theoretical inferences, Heidegger's approach reflects its relationship to American pragmatism by extending assertion to the individual's everyday activities within the physical world.

For Heidegger, a useful thing is "essentially something in order to..." such as serviceability, helpfulness, usability, handiness, [which] constitute a totality of useful things" (p. 64). He argues that the meaning of "in order to" includes a "reference of something to something" (p. 64; emphasis in original). In particular, things are seen as useful in terms of their belonging to other useful things" (p. 64):

"... writing materials, pen, ink, paper, desk bloter, table, lamp, furniture, windows, doors, room[:]. These "things" never show themselves initially by themselves, in order then to fill out a room as a sum of real things. What we encounter as nearest to us, although we do not grasp it thematically, is the room, not as what is "between the four walls" in a geometrical, spatial sense, but rather as material for living. (p. 64)"

Thus the notion of handiness cannot be grasped or explained theoretically; rather, it can only be apprehended as an association that "makes use of things" and "guides our operations and gives them their specific thingly quality" (p. 65). For example, a hammer is associated with hammering; but the act of hammering itself "discovers the specific 'handiness' of the hammer" (p. 65). For Heidegger, then, the handiness of a useful thing is discovered in its association with the "work to be produced" (p. 65):

"The shoe to be produced is for wearing (footgear), the clock is made for telling time. The work which we primarily encounter when we deal with things and take care of them—what we are at work with—always already lets us encounter the what-for of its usability which essentially belongs to it. (p. 65; emphasis in original)"

Within the framework of Heidegger's notion of handiness, basic level categories are "useful things" that "reveal" themselves as "things" by their handiness. Moreover, it is possible that it is exactly this "thingly" (p. 65) quality that underlies the predominance of basic level categories in everyday vocabulary. While the predominance of basic level categories may be due, in part, to a general understanding of "work to be produced" (p. 65), Heidegger sees the work produced not only as a reference to the "what-for of its usability" of a thing and the "whereof of which it consists" (p. 66), but also to the "surrounding world of nature" (p. 67). He provides as an example the use of a clock, arguing that:

When we look at the clock, we tacitly use the "position of the sun" according to which the official astronomical regulation of time is carried out. The surrounding world of nature is also at hand in the usage of clock equipment which is at first inconspicuously at hand. (p. 67)

For Heidegger, the handiness of things in the "work world nearest to us" (p. 67) is based on the function of discovery and depends upon the "way we are absorbed[:] innerworldly beings that are brought along together with their constitutive references are discoverable in varying degrees of explicitness and with a varying [degree of] attentive penetration" (p. 67). In other words, the use of "things" is directly related to the individual's perception and understanding of those things and, as such, those things do not have an independent meaning that exists outside the individual's understanding. As Dewey argued (1997), the dependence on human perception and understanding to ascribe meaning to a thing represents the knowledge that enables individuals to "adapt [the] environment to our needs ... [and] aims and desires to the situation in which we live" (p. 217). Thus, for Dewey, knowledge is comprised not only of what we are now conscious of" but also of the "dispositions we consciously use in understanding what now happens" (p. 217).

Within the framework of Heidegger's notion of handiness, variation is both natural and phenomenological in that perception and understanding -- and thus the meaning of "things" --arise out of the individual's contextualized experiences of engaging with objects. Basic
level categories vary across individuals and across cultures because of differences in the everyday experiences and activities of individuals. The implications of handiness for basic level category research in LIS is obvious: The operationalization of basic level categories based on the hierarchical structure of normalized vocabularies cannot account for variations in the relational structures of individuals, much less variations in shared conceptual structures across cultures. A more effective approach to operationalizing basic level categories must take into account the contextualized experiences of individuals actively engaging with objects in order to assess their understanding of an object's "thingly quality" (p. 65) and its relationship to other objects in the everyday world of work.

References


Hoenkamp, E., Stegeman, O., and Schomaker, L. 1999. Supporting content retrieval from WWW via "basic level categories.” *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, USA*, 311-312.


Rorty, R. Essays on Heidegger and others: Philosophical papers. Cambridge: Cambridge University Press
Epistemological Basis of some Common Categories - 
A Study of Space and Time as Common Concepts

Abstract
Common categories (often termed as Common Subdivisions) have been an integral part of all Library Classification Systems. The number of common categories, applicable to all classes of subjects stood at Seven (7) tables of common subdivisions and among these Time and Space categories manifest both as common and special categories. Common categories in Library classification was extensively examined by de Grolier, and Ranganathan has also suggested there is scope for comparative study. Time is also treated in dissimilar manner in different classification systems. The paper presents some unified approaches to space and time categories and in this context seeks their epistemological basis.

Introduction
The study and implications of common categories in Library classification has been extensively examined by de Grolier (de Grolier, 1962) The Classification Research Group also made attempts to evolve general categories. Common categories are also labelled as Common Subdivisions; they occur as facets of a subject and also as categories in specific domain of knowledge. Identification and application of categories in indexing can be traced to the “Concrete and Process” concepts enunciated by Julius Otto Kaiser. Similar conceptualisation was made in automatic indexing by others and in a recent paper by Giunchiglia and Dutta on digital library (Giunchiglia and Dutta, 2011). Ranganathan in his words termed it “to determine and maintain the scale of preferred neighbourhood relations among all ideas and among all subjects (Ranganthan, 1967, 395).

Duyvis observed there is no uniform approach to treat them (common subdivisions) in Library Classification and suggested that there is need for research (Duyvis, 1955). Hence an extensive comparative study of their treatment in general classification systems was undertaken (Asundi, 1982). The Tables of Common Subdivisions in general classification systems – both active and non-active and in some special classification systems - were compared. The final study selected 7 Tables of Common Subdivisions, in the 4 active general classification systems and noticed incompatibility among them and suggested a unified scheme of common subdivisions to remove the incompatibilities. In this context the study sought to use the epistemological basis for three of the seven Common Subdivisions, as they are distinct from others, and also as domain specific common categories. For this purpose, in library classification some ideas often are ordered in conjunction, because of their mutual intrinsic relation as with „space and time“. It is desirable to consider the epistemology of knowledge entities so that a unified structure is framed and adopted in knowledge organization.

Space and time also need consideration to understand their intrinsic characteristics. This is possible through the study of their epistemological foundations – the typology of relations that exist in the knowledge of space and time individually. The status of Space and Time makes them universal entities, applicable to all subjects. Space and time at their verbal representation also need equal attention in understanding their core/intrinsic characteristics. In most classification systems, because of their mutual relation and occurrence these are arranged / treated together. The knowledge structure of space and time has to be elaborated consequent to their universal application not only in library classification but in knowledge organization in general. In this paper, an attempt is made to examine the epistemological basis of Space and Time as knowledge entities and to evolve
their knowledge constituents that would help to unify their varying treatment in different Classification systems.

**Union of Space – Time**

Spatio-temporal idealism is often related to space and time frame of thinking. The most common and conjectured use of space and time as knowledge entities is found in all Library Classification - knowledge and document classification systems. The classification of Space per se is often distinguished; and that has led to its anomalous treatment in different classification systems rather than as a concept of knowledge. Ranganathan has stated that through the comparative study of space in some classification systems it is possible to culture (*acculturation*) this area of classificatory discipline falling within the purview of the categories time and space. Time the inseparable idea from space is also treated in dissimilar manner in classification systems. Most of them conform to the Principle of Later in Time, i.e. from Ancient to modern or vice versa. This linear pattern does not match with the multi-dimensional knowledge of time as revealed from its epistemology. There is also a case for the study of time concept from its epistemological foundations to configure its multidimensional nature and structure.

The most general aspect of space emerges from the earliest concept of “cosmological world”. The philosophical content of „Cosmology” has been both space and time – that includes finite and infinite (limitless) world. The finite quantity might be related to the earth and the limitless to the space and beyond earth. Though space and time tend to co-exist, and often quoted in mutual relationship, there are arguments, such as whether the beginning of universe had in time or the universe was in existence before „time.” The latter is termed as empty time. This method has also been later followed by philosophers and classificationists though they have diverse views. The growth of knowledge process has been elicited here also for space and time.

**Epistemological Basis of Space and Time**

Ranganathan stated space as a uni-dimensional entity, so it has whole-part and general subject aspect of relation. He has also suggested a need for comparative study of space (Ranganathan, 1967, 399). The most general aspect of space epistemologically emerges from the earliest concept of cosmological world. The philosophical content of cosmology has been both space and time, which includes both finite and limitless world. Of late the “space” is collective to be as the (physical) form of a world, which is finite with limitless bounds (beyond earth). The finite quantity might be assumed to constitute the earth and the limitless outer space – beyond earth. Considering the beginning, space as cosmology, the pattern of space subdivisions in classification would start from celestial to terrestrial sequence of organization. In the next order it can follow from the terrestrial to micro-level administrative units of political organizations as may be found in modern structure. In philosophical treatises space and time also exist together and are often treated together. In summary they also coexist from a situation beginning with universe, however doubts are raised in fair arguments whether the beginning of the universe had in time or the universe was in existence before time. In the latter case it was termed “empty time”. In recent periods time has received a universal appraisal with Unesco Expert Meeting to suggest a study on the expression of time in different domains. In this context Neelmeghan and Narayanan made a study of differences in concept, measurement, impact and expressions of and about time in different cultures (Neelmeghan and Narayana, 2012). So the dimensions of space and time have attracted the attention of knowledge organisation systems since ancient times.
Epistemological Basis of Space

The cognition of space has been different in different periods of human civilization. Its demarcation derived also varied from time to time with different areas of comprehension. The shape of earth as flat, was a firm belief even in the 16th Century A.D. and so the assumption of space as different in different periods. But the elucidation attempted here are those conceived ideas of space that were in the time past, from the ancient periods. The Greek era correlated it to cosmic space, to universe, theology and physics. The present context of space in contrast is rather in the last percept as an entity of physics, and of exploration of astronomical bodies. But this does not mean that it has changed its earlier concept of cosmic universe as its ancient value is still extant. Since the period of Greek much has been explored and achieved in revealing the secrets of space. The space in its earliest summation was rather deceptive to modern thinking as compared to its ancient, logical and ontological perception. Such idea of space has unlimited expansion and it could be infinite space when space is not limited by matter. But the limitations with matter bring it to the finite space like earth and other planetary bodies.

The modern concept of space has come to be understood since 1600 A.D. John Locke has dealt space as matter, and thus earth aspect as matter in the broad meaning brought it in close relation to space. Kant also brought to light earth space relation. Thus space comprises all descriptive elements which relatively coexist since ancient to modern times (Dictionary, 1973, 295).

Historically various status of space since Greek period may be categorized into the following broad divisions.

1. Makon - From Greek Sources
2. Chaos - Aristotle refers; Heroids class in space
3. Cosmos - Or Ordered Universe and world order as suggested by Homer
4. Earth - Space and matter relation by Kant and Leibnitz
5. Space - Matter relation by John Locke

Considering all these broad descriptive elements, a set of representations of space in knowledge domain with the following structure is derived. The concept of space (isolate) in knowledge organization may be represented as an insolate comprising of broad types of space as follows:

- Celestial;
- Terrestrial;
- Fossilic;
- Physiographic;
- Socio-political; and
- Politico-economic

Under each group appropriate subdivisions are possible. For example, groups like EEC, OECD, SAARC would be part of the „polito-economic” group. Any later developments are only possible through extrapolation.

Epistemological Basis of Time

Aristotle’s contention of time as „continuous quantity” is measurable, as past, present and future. Another attribute inherent in the above statement is continuity which implies it is divisible nature. Similarly other quantities of time are also subjected to division. Hawking (Hawking, 1988) has stated Immanuel Kant’s influence in relation to classification of time gives an attribute; it is yet another way. In case of time the reasoning supported by him is a valid a priori. It is a form of inner and indirectly in outer sense, empirically real and
transcendently ideal condition. His assertion is that time is a necessary condition of all experiences, both introspective and sensory. So also the idea of time as a component of classification expands these characteristics (Dictionary, 1973, 389-406).

Time, manifests empirical and transcendental attributes in Microtime, Macrotime and both are being real and basic units of measurement. The other two - historical periods and ancient times are ascribed to as intellectual and intangible aspects. The latter two contain socio-cultural ideologies of the past which for the present are imaginary for the intellect and thus intangible. The Microtime and Macrotime constitute units of measurement of time in all its manifestations, therefore they are basic and real manifold. Thus based on these logical reasoning, the time as period in a larger dimension is grouped into four structured ideas which further manifest as “Absolute, Relative and Recurrent”. This order of time is universal and holds good in all situations and conceiving it comprises all subdivisions of period. To test this hypothesis, the comprehensiveness of subdivisions expected to encompass in these groups, the Mach’s Thesaurus dictionary (82) and Laffal’s Concept Dictionary of English are compared (Laffal, 1973, 279-280). Most of the concepts it is found enumerated in these sources in relation to time and which would form the subdivisions under the unified structure given in the chart below. The definitions of the concepts of time are also presented.

The charting of Microtime, Macrotime as attributes and Absolute, Relative and Recurrent as characteristics results in identifying all the knowledge components of time.

Table 1: Unified Knowledge Structure of Time

<table>
<thead>
<tr>
<th>Characterization Manifestation</th>
<th>Absolute Time/ Definite Demarcation/ Portion of time</th>
<th>Relative Time/ Time in relation to particular period</th>
<th>Recurrent Time/ Frequency of time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attributes/Dimensions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Microtime</td>
<td>Seconds</td>
<td>Morning/ Noon</td>
<td>Daily, Weekly</td>
</tr>
<tr>
<td></td>
<td>Minutes</td>
<td>Evening, Night</td>
<td>Fortnightly,</td>
</tr>
<tr>
<td></td>
<td>Hours</td>
<td>Midnight</td>
<td>Monthly</td>
</tr>
<tr>
<td>Macrot ime</td>
<td>Day</td>
<td>New Moon</td>
<td>Yearly</td>
</tr>
<tr>
<td></td>
<td>Week</td>
<td>Seasons</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fortnight</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Month</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Historical Periods</td>
<td>Luster / Lustum</td>
<td></td>
<td>Anniversary</td>
</tr>
<tr>
<td></td>
<td>Decade</td>
<td></td>
<td>Decennial</td>
</tr>
<tr>
<td></td>
<td>Quinquennium</td>
<td></td>
<td>Jubilees</td>
</tr>
<tr>
<td></td>
<td>Life Time</td>
<td></td>
<td>Centennial</td>
</tr>
<tr>
<td></td>
<td>(60 Years)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ancient Times</td>
<td>Generations</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Era</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Epoch</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Millennium</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusion

Space and Time are universal entities of Library Classification since its genesis. The paper demonstrates the unified scheme of Space, Time developed using their epistemological basis. This scheme has potential to perpetuate, for interpolating all future elements and concepts of the selected concepts of common categories – space, time. I have also used and applied several conceptual bases, including the categories and relations suggested by Kaiser in developing the unified scheme for other common subdivisions.
References
Note: This Part – 1 of Series of 7 papers being published in “*Information Studies*”. The Part-2 of the series was published in Volume 18(2): 2012.
Domain Specific Categories and Relations and their Potential Applications: A Case Study of Two Arrays of Agriculture Schedule of Colon Classification

Abstract

The categories/isolates are broadly conceived as common and special. The common categories are applicable to all the classes of subjects in a Classification system, whereas the specials are applicable within a domain or specified classes of a classification system. The CC has represented some unique special categories, especially in the Agriculture Subject schedule, and such a provision is not seen in any other classification system; not even in any other subject schedule of Colon Classification. These special categories are termed here as “Domain Specific Categories”. The paper analyses the thematic relationships within and outside the subject schedule with potential applications in devising a scheme of metadata as demonstrated in a research study on Indian Medicinal Plants. The other potential applications of such thematic relationships are in the creation of semantic maps and in linking concepts from different domains of knowledge.

Introduction

The Universe of Knowledge can be segmented as, known, unknown and knowable. The concept of domain connotes known knowledge, representing several disciplines, the fields of study which in other words are the cognitive content of known knowledge. The universe of knowledge, and its one segment, known-knowledge include (See Fig.1):

![Fig. 1: Derivation of Isolates(Categories)](image)

A lateral relation is visualised between universe of isolates and universe of subjects. Many subjects consist of some domain specific categories.

Studies on Common and Special Categories

Categories have been an integral part of every library classification system, especially devised and developed since the Dewey Decimal Classification System. The study on the Common Subdivisions in General Classification Systems undertaken by the author underlines this. (Asundi, 1982). The general categories or common isolates as termed by Ranganathan and their relations in knowledge organisation are found in three levels of application - as common subdivisions in library classification systems, secondly as the facets of a subject – the Five Fundamental Categories of Colon Classification (CC), and thirdly in indexing as general categories to regulate the citation order. The concept of domain specific categories, their relations and the potential applications outside Library Classification has been demonstrated by the author (Asundi, 2001). In this context the following statement by Deokattey and others (Deokattey et al, 2010. 175.) clearly substantiates this claim.
A domain ontology is a knowledge organisation tool on a specific subject domain either pure, interdisciplinary or multidisciplinary and which incorporates both a metadata schema and a vocabulary.

In the subject schedules of any classification scheme, there are some inherent domain specific categories with potential to represent knowledge within and outside their subject purview. For instance *Music* as a basic subject has two main modes of presentation – *Vocal* and *Instrumental*. The Musical Instruments though numerous would be brought under a finite number of categories – as *Keyboard, Wind, Percussion and String* instruments. Similarly, in the organisation of Paintings a representation would be made by a finite number of concepts like- *Forms and subjects; Material/Foundation/Surfaces; Techniques and Procedures; Styles and Schools;*, and finally by *Periods and Phases*. A study on verbal representation of paintings was made by the author (Asundi, 1996). These finite number of categories found in some domains of knowledge are commonly found in most classification systems, and could be termed as domain specific common categories. These can find application not only in the Classification of books but also as knowledge representation tools in digital libraries and in the classification and categorisation of museum objects.

Neelameghan and Prasad have shown that mapping the concepts of subject areas can play complementary role between classification and thesauri, and this relationship is useful in designing user interfaces and in many other areas (Neelameghan and Prasad, 2001, 97). In the context of establishing lateral relationships among domains, Neelameghan and Raghavan have presented the “Semiotics of relationships in knowledge organisation (Neelameghan and Raghavan, 2006, 117-118). The above have clearly demonstrated that the “domain specific categories” can lead to the process of establishing relationship within and outside a domain, and have implied uses in devising metadata, designing user interface, creation of thesaurus and searching information systems.

**Objective and Purpose of the Paper**

The two Arrays of classes of Agriculture Schedule of *CC* have remained conspicuously unnoticed since their introduction in the 3rd Edition of *CC* based on the research work undertaken by Krishna Rao. Among knowledge organisation systems, ontologies and concept maps play an important role. These are inherent in this scheme of classes and have remained unexplored so far.

**The Case Study of Agriculture Schedule**

It is from the studies of several classification systems, the two Arrays of Agriculture Schedule of *CC* have some uniqueness which is an exception to any general or special classification systems. The unique type of domain specific common categories for Agriculture were evolved out of extensive research conducted by Krishna and have evolved to the present state in *CC*. This has been applied in the design of a Metadata Schema for the Indian Medicinal Plants (Asundi, 2001). The paper presents the unique features and the potential application of domain specific common categories in the Agriculture schedule in *CC* 7th Edition and their application in knowledge organisation in the digital library environment. The two arrays enumerated in the 7th Edition of *CC* are presented here for ready reference (Ranganathan and Gopinath, 1987, 90) (see Table).

**Classification of Crops in Agriculture Schedule**

In all library classification systems the crops as agriculture produce are categorised as; Specific plant crops – further divided as; Field and Plantation Crops, Orchards, fruits, forestry, Garden Crops, Edible Tubes and Bulbs and so on. This almost requires enumeration of all Crops and may lead to overlaps, e.g., Plantation crops and Forestry and Garden Crops and others. *CC* being a Freely Faceted Classification has overcome this anomaly by dividing all agriculture produce in two Arrays – Utility Array and Organ (Part) Array.
Table 1: Array of Isolates of Order 1 and Order 2 in J – Agriculture

<table>
<thead>
<tr>
<th>Array of Order 1 – By Utility</th>
<th>Array of Order 2 – By Organ/Part</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decoration</td>
<td>Sap</td>
</tr>
<tr>
<td>Feed</td>
<td>Bulb</td>
</tr>
<tr>
<td>Food</td>
<td>Root</td>
</tr>
<tr>
<td>Beverage</td>
<td>Stem</td>
</tr>
<tr>
<td>Stimulant</td>
<td>Leaf</td>
</tr>
<tr>
<td>Oil</td>
<td>Flower</td>
</tr>
<tr>
<td>Drug</td>
<td>Fruit</td>
</tr>
<tr>
<td>Fabric</td>
<td>Seed</td>
</tr>
<tr>
<td>Dye, Tan</td>
<td>Whole Plant</td>
</tr>
<tr>
<td>Adhesive</td>
<td></td>
</tr>
<tr>
<td>Manure</td>
<td></td>
</tr>
<tr>
<td>Vegetable</td>
<td></td>
</tr>
</tbody>
</table>

Sugar Producing

The Utility Array is based on the use of the produce as Food, Beverage, and Stimulant Adhesive, and so on and comprehensively comprises all possible utilities of the agriculture produce. The Organ/Part Array consists of the parts of the plant used as; food, beverage or stimulant or an adhesive, and logically the organs (Parts) of the plant are arranged as per the Principle of Bottom Upwards. Another universal concept adopted by the Agriculture schedule in the representation of the crops, is not by the order of “favoured country” as is done in Dewey Decimal Classification (DDC) \(^1\), but the more universal favoured category of food grain by its utility.

Illustrations

Taking *RICE* as an example its Semantic Map is presented below.

*Fig. 2: Semantic Map of RICE as an Illustration*

- **Utility Array Isolate** – Food
- **Organ Array Isolate** – Seed

The most consumed Food grain in the world (Universally favoured)

Concurrently the Class Number for Rice as per the Ed7 of CC is “J381” and is graphically analysed below to represent the above semantic map

*Fig. 3: Analysis of Class Number for RICE as per Colon Classification*

<table>
<thead>
<tr>
<th>Basic Subject</th>
<th>Utility Array Isolate</th>
<th>Organ Array Isolate</th>
<th>Favoured Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGRICULTURE</td>
<td>FOOD</td>
<td>SEED</td>
<td>1</td>
</tr>
</tbody>
</table>

Potentially this conceptualisation of domain specific common categories as found in the Agriculture schedule of CC enables construction of Class Numbers for almost all agricultural crops including field crops, plantation crops, forestry produce, etc. This is the uniqueness of the two Arrays of Agriculture schedule of CC.

Knowledge Organisation Components

Knowledge organisation systems such as ontology, semantic maps and taxonomies are also embedded in this schedule. The Utility isolates, termed as Quasi isolates by

\(^1\) Note: In DDC almost all special isolates of a class begin with special isolates pertaining to “USA”. E.g. 811 American Literature, which is termed here as “favoured country”
Ranganathan represent ontologies and the Entities like Rice can be presented as a Semantic Map as demonstrated above. The taxonomical conceptualisation is found in the categorisation of the crops by the geographical and field orientations would also find appropriate representations.

An example of a Semantic Map constructed for Coconut Tree as an illustration is presented below. In India, the Coconut Tree is conceived as “Kalpavriksha” a tree of many utilities; in the sense that every part of the tree has utility.

**Fig. 4: Showing multiple links using the two Arrays**

<table>
<thead>
<tr>
<th>Food</th>
<th>Oil</th>
<th>Fruit Core (Inner Part)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coconut Tree</td>
<td>Utility</td>
<td></td>
</tr>
<tr>
<td>Part (Organ(s)) Array – Fruit/Nut</td>
<td>Shell (Outer part)</td>
<td></td>
</tr>
<tr>
<td>Agro-based Industry</td>
<td>Coir Industry</td>
<td></td>
</tr>
</tbody>
</table>

**Metadata Schema for Medicinal Plants**

This Semantic Map presents how the two arrays – Utility and Part (organ) can be viewed linking concepts within and outside the farming domain to industry domain (shown with dotted line). Similar applications are made in evolving the Metadata elements for the Indian Medicinal Plants and their parts utility in the context of drug or medicine as described below.

It is conceived that, the main sources of Medicine in the Indian System of Medicine – the Ayurveda are plants. The medicinal value is found in one of the parts of the plant or the whole plant. The Utility Array has an Isolate – Drug and the Organ Array comprises all parts of the plant and also represents the whole plant. A plant having a medicinal value gets the element “Drug” as its metadata element and then what organ of the plant has the medicinal utility – for instance turmeric is a root/bulb and the root or bulb is it’s other metadata element. So the two Arrays – Utility and Organ have found application in deriving metadata elements for Indian Medicinal Plants. More empirical studies of the kind need to be carried out.

**Conclusion**

The paper has demonstrated the potential value and utility of domain specific categories. It is not exhaustive covering all domain specific categories, not only in CC but in other general classification systems, for instance the once used “Centred Headings” (now Add as) in the Dewey Decimal Classification which also needs to be researched. A study of Area Tables of Dewey Decimal Classification from Editions 11 to Edition 19 to show the impact of social, political, economical and ethnic characteristics in the changes of the domains and boundaries of the countries of the world has been reported (Asundi, 1976). The changes can be compiled and can form a gazetteer- like tool for knowledge organisation. So the potential of domain specific categories have to be explored for design of tools for knowledge organisation.
References and Notes

Asundi, A.Y. 1976. An analytical study of three schedules of Dewey Decimal Classification – A study in Commemoration of Dewey Decimal Classification Centenary Year. *A project work in partial fulfillment of the Training Course in Documentation and Reprography (1975-76)*. Indian National Documentation Centre, New Delhi, India.


K.S. Raghavan – DRTC, Indian Statistical Institute, Bangalore, India
A. Neelameghan – Sarada Ranganathan Endowment for Library Science, Bangalore, India

Indic Cultures and Concepts:
Implications for Knowledge Organization

Abstract
This paper deals with the impact of culture, especially indigenous cultures, on conceptualization and semantic relationships. Representation of concepts in the performing arts, mythology, humanities, and cognate domains viewed through different cultural perspectives in knowledge organization and knowledge organization tools (KOTs) is examined.

Introduction
We live in a culturally diverse world. Cultural differences among communities have come to the fore in recent years thanks to developments in information and communication technologies (ICT) leading to increased interaction among peoples of different cultures. Depending on the context, „Culture“ may signify: (1) a particular society at a particular time and place, e.g. Vedic culture, Dravidian culture, Mayan culture; or (2) the knowledge and values shared by a society / community, e.g. modern Bengali society; or (3) the tastes, styles in art and manners that are favoured by a community or social group, e.g. Elizabethan poetry, French Renaissance art, and (4) the attitudes and behaviour that are characteristic of a particular social group or organization, e.g. Jain monks, Parsis of India, South-Pacific Islanders, etc. Culture and the environment in which a culture evolves have always been major factors influencing knowledge and concept formation, development of language and terminology and formation of links between concepts among the members of a cultural group. In other words the nature of concepts, the terminology used to represent concepts, the range and kinds of interrelationships between concepts are all functions of the culture and social life of the community. The culture and social life in turn are impacted and determined by a variety of factors including the geographical terrain in which the culture evolves, the dominant flora and fauna in the terrain, the interactions between the members of the cultural group and other cultural groups, the predominant occupations of the members of the cultural group, etc. A cultural frame could be seen as a perspective bound by certain traditions that would relate to a particular culture. For example, the picture of someone wearing a particular dress may state something about the cultural and time frame. If required, the term „culture“ may be used to mean a country, community, linguistic or ethnic group. It is precisely this that makes us speak of Indic culture, Islamic culture, Tamil culture, etc. This manner of understanding culture is possibly a consequence of the marked characteristics and traits of our societies – regions, linguistic groups, religious groups, etc. All these have major implications for the development of tools for knowledge organization (KOTs) for culture-specific domains as concepts, their names (terms), semantic inter-relationships between them and organization in meaningful ways are basic to both knowledge formation and the development of KOTs such as classification schemes, thesauri, taxonomies, ontologies, etc. KOTs need to accommodate and display such different cultural viewpoints and relationships among concepts. A concept is a mental representation that stands collectively for all meaningful statements that can be made about an entity. All the meaningful statements that can be made about the flower *lotus* constitute the concept of *lotus*. It should be obvious that the concept of *lotus* as viewed from the
perspective of a particular linguistic and cultural group will not be the same as seen from the perspective of another group. For example, certain South Pacific islanders are reported to be knowledgeable about some twenty-five different varieties of a single fish species, including attributes such as when and where they will be seen in the sea, which ones are edible or usable for medicinal purposes, etc. Such an elaborate categorization of one type of fish may not be found in text books on fish. However, such an extensive categorization is important to the daily livelihood of the community concerned. Similarly, a flower named "kuRinci" appears to have some significance in the life and culture of Tamils belonging to the Dravidian culture and civilization. There are several different aspects to the flower – cultural, religious, scientific, literary, mythological, etc and literature exists on all these different aspects of the flower. The term is also used in Tamil to denote a number of related concepts. For example, the term has also been used by early Tamils to denote the kind of terrain similar to those in which the flower grew in abundance. Over a period of time cultural aspects such as music, poetry, etc also appear to have been denoted by the same term. Thus the Tamil term kuRinci meant several related things and its connotation has therefore to be understood depending on the context in which it is used.

**Objectives of the Paper**

The need for and importance of effective and adequate representation of culture-specific concepts in knowledge organization tools and facilitating navigation among related concepts cannot be overemphasized. Information retrieval systems have traditionally used thesauri and similar KOTS for representing concepts as also for displaying semantic relations between concepts. Most such KOTS have limitations in terms of both:

- The adequacy of coverage of concepts encountered in discourses; Most KOTS evolve and develop in a certain environment. While there may be a considerable degree of agreement on the nature, naming and connotation of concepts in the sciences and engineering, this is not so in the social sciences and even less so in the humanities and culture-specific domains. For example, a thesaurus for the domain of Education or Sociology developed for an information system in the U.S. or Western Europe may not be adequate in many respects when it comes to indexing literature on the domain originating from say, India or China; i.e. there could be several concepts and conceptual relations unique to the domains of Education or Sociology as perceived in the Indian or Chinese contexts for which there may be no corresponding or equivalent concepts in the western cultures;

- The thesauri that have been in use in retrieval systems group all semantic relations into a limited number of categories. The three most widely used categories of relations in thesauri are Equivalence, Hierarchical and Non-Hierarchical Associative Relations. While there is a certain degree of agreement on what constitutes a Hierarchical relation, a wide range of relations obtaining between concepts are treated as Non-Hierarchical Associative Relations (Lateral Relations). In practice this makes navigation between semantically related concepts difficult as the end user is unable to differentiate between different kinds of Lateral Relations. The availability of more powerful tools for building information systems and supporting navigation (e.g. Ontologies) has made it worthwhile to explore the feasibility of categorizing lateral relations into a limited number of categories and representing these.

In recent years a few projects related to developing KOTS required to support design of multilingual databases, digital collections and related user interfaces in culture-specific domains have been initiated. This paper is largely a result of the experiences gained while working on these projects. The objective of this paper is to examine the impact of culture, especially ancient traditions and indigenous cultures, on conceptualization and semantic
relationships. Two closely related issues that have implications for knowledge organization will be examined in this paper:

- The nature of concepts encountered in culture-specific domains; the impact of culture on concepts formation will be specifically examined. This will be done using a few examples representative of the nature of problems encountered.
- The nature and kinds of semantic relations encountered between concepts and their representation: More specifically the kinds and range of non-hierarchical associative relations between concepts in culture-specific domains will be examined.

Concepts and Conceptual Relations in Indic Cultures

Several major civilizations and religions have had their origins in India. Because of this India has a rich philosophical heritage right from the Vedic-Upanishadic period to the scholastic period; and many schools and sub-schools of philosophical thought have emerged. „Hinduism“ connotes a way of life rather than a religion. The more orthodox followers may prefer to use the term „sanatana dharma“ meaning the „ancient and eternal religion” instead of Hinduism. Looked at from this angle, all religions of Indian origin – Jainism, Buddhism, Sikhism are different facets of Hinduism. However, in practice, the term is applied specifically to the religion based on the Vedas and the Vedic philosophy. The basic scriptures, the sources that provide the overall framework for the interpretation of knowledge/concepts and philosophical thoughts of Hinduism, are the Vedas. As against this Vedic culture, the civilization of the ancient speakers of the proto-Dravidian languages is broadly referred to as Dravidian civilization. The history of this is obscure and there is more than one theory about their origin. One theory is that even before the arrival of the Aryans in the Indian subcontinent, the speakers of the proto-Dravidian languages (comprising Proto-North Dravidian, Proto-Central Dravidian, and Proto-South Dravidian) had settled in most parts of the subcontinent. The foreign invaders forced the Dravidians, the original inhabitants of India to the southern parts. In this paper we examine the nature of Indic culture-specific concepts and their interrelations with emphasis on the Dravidian culture and civilization. The Dravidian culture is distinct from the Vedic culture even though because of extensive and continuous interactions between the two over thousands of years, a composite culture has evolved. However, in handling literature on arts, architecture, Music, philosophy and many other culture-specific domains, it is imperative to handle and accommodate concepts and conceptual relations that are unique to the Dravidian culture. In the following paragraphs, we illustrate, with a few examples, the nature of concepts and conceptual relations in Dravidian culture-specific areas and also examine briefly the problems and issues related to developing KOTS for design of information systems in such domains. Most regions of the southern states of India are home to the Dravidian culture. The region has unique art forms such as dance and music, folk forms, architecture, festivals, rituals, fauna and even calendars distinct from those of the rest of the country. These art forms have also spread to some parts of Southeast Asia and many of the elements of Dravidian culture can be found, for example, in the art forms and architecture of Cambodia and Indonesia.

The Carnatic Music System

The classical music system that is prevalent in most parts of South India is known as Carnatic music. As early as between 200 BCE and 1200 CE, long before systematization of this form of music into a discipline, systems of music and dance are said to have been developed, e.g. by the Tamils. There are references to this in many works of the sangam period. For example, there are references to a highly evolved form of music among the Tamils in such works as tolkappiyam (a Tamil grammar work of the Sangam period).
classification of the terrain (land) into five categories viz., mullai (pastoral), marutham (arable land), kurinji (hilly areas), paalai (waste land), and neytal (coastal) became the basis for classification of ragas (melodies). They developed pann (ragas) of their own and
the panns were categorized based on the appropriate time of the day for singing the raga:

- pakalpanns (day time melodies)
- iravupanns (night-time melodies)
- potupanns (melodies that could be sung any time of the day / night)

The present prevalent form of Carnatic music has as its basis, the system of ragas (melodic scales) and talas (rhythmic cycles). It is highly structured and there are seven rhythmic cycles and 72 fundamental (Melakarta) ragas. All the other ragas originate from these. A characteristic feature of this form of music is the highly devotional element of most of the compositions that are generally sung / played during a concert. An idea of
the web of lateral relations between a music composition and entities associated with it can be obtained from the following table which associates the navagraha compositions of Muthuswami Dikshitar – one of the trinities of carnatic music.

Table 4: Entities associated with a music composition

<table>
<thead>
<tr>
<th>Graha / Planet</th>
<th>Symbol</th>
<th>Deities</th>
<th>Foods offerings</th>
<th>Element</th>
<th>Colour</th>
<th>Gemstone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sun</td>
<td>Surya</td>
<td>Rudra / Agni</td>
<td>Wheat / Sugar</td>
<td>Fire</td>
<td>White</td>
<td>Ruby</td>
</tr>
<tr>
<td>Moon</td>
<td>Chandra</td>
<td>Gautri</td>
<td>Rice</td>
<td>Water</td>
<td>White</td>
<td>Pearl</td>
</tr>
<tr>
<td>Mars</td>
<td>Angaraka</td>
<td>Ksetrapati / Prthvi</td>
<td>Peas</td>
<td>Fire</td>
<td>Red</td>
<td>Coral</td>
</tr>
<tr>
<td>Mercury</td>
<td>Budha</td>
<td>Narayana / Vishnu</td>
<td>Green lentil</td>
<td>Earth</td>
<td>Yellow</td>
<td>Emerald</td>
</tr>
<tr>
<td>Jupiter</td>
<td>Brhaspati</td>
<td>Brahma / Indra</td>
<td>Peanuts, rice, curd</td>
<td>---</td>
<td>Yellow</td>
<td>Yellow sapphire</td>
</tr>
<tr>
<td>Venus</td>
<td>Sukra</td>
<td>Marvat / Indrani</td>
<td>Small beans, ghee</td>
<td>water</td>
<td>White</td>
<td>Diamond</td>
</tr>
<tr>
<td>Saturn</td>
<td>Sani</td>
<td>Prajaaspati</td>
<td>Sesame seeds &amp; oil</td>
<td>space</td>
<td>Blue</td>
<td>Blue sapphire</td>
</tr>
<tr>
<td>Dragon’s Head</td>
<td>Raahu</td>
<td>Sarpapara Jaya / Godhaavati</td>
<td>Black lentil</td>
<td>---</td>
<td>Smoke</td>
<td>Hessonite / Garnet</td>
</tr>
<tr>
<td>Dragon’s Tail</td>
<td>Ketu</td>
<td>Brahma / Citragupta</td>
<td>Horse grain, Lemon rice</td>
<td>---</td>
<td>Smoke</td>
<td>Cat’s eye</td>
</tr>
</tbody>
</table>

Musical instruments are also associated with deities; e.g. veena is associated with goddess saraswati, flute with lord Krishna, and so on. There is also literature on the suitable kind of wood for making certain musical instruments; certain kinds of melodies are associated with certain rituals. All these suggest extensive interconnection between music and folklore, religion, philosophy, mysticism, other performing art forms, sociology, festivals, customs, etc.

Flora and Tamil Culture

A bell shaped flower, kurinji, grows in the Kodaikanal and Nilgiris hills of South India. The flower is of interest to botanists as it blossoms only once in twelve years unlike the other common flowers. An examination of the literature on the flower suggests that there are several different aspects to it – cultural, religious, scientific, literary, and mythological; literature on all these different aspects exists. The term is also used in Tamil to denote a number of related concepts and the semantics of the term changes depending on the context in which it is used. For example, the term kurinji has also been used by early Tamils to denote the kind of terrain similar to those in which the flowers grew in abundance. The early cultures appear to have emerged in environments that were conducive for living and the areas were identified by the flora that grew in abundance in those regions. Over a period
of time other cultural aspects such as music and poetry also appear to have been denoted by the same term. An idea of the web of relations between the term and other concepts can be obtained from the table below.

Table 2: Homographs in Tamil

<table>
<thead>
<tr>
<th>Term</th>
<th>Flower</th>
<th>Terrain</th>
<th>Music</th>
<th>Poem</th>
<th>Poetic convention</th>
<th>Deity</th>
</tr>
</thead>
<tbody>
<tr>
<td>kuRinji</td>
<td>kuRinji</td>
<td>Hilly Terrain</td>
<td>Music associated with hilly terrains</td>
<td>kuRinci-p-paaTTu</td>
<td>Clandestine union of lovers associated with hill tract</td>
<td>kuRinji Andavar</td>
</tr>
<tr>
<td>mullai</td>
<td>jasmine</td>
<td>Forest tract</td>
<td>Melody type associated with forest</td>
<td>Mullai-p-paaTTu</td>
<td>Patient endurance of lady during separation from lover, associated with forest land</td>
<td></td>
</tr>
<tr>
<td>marutam</td>
<td>Flowering murdah</td>
<td>Agricultura l tract</td>
<td>Morning melody special to agricultural land</td>
<td>Love action associated with agricultural land</td>
<td></td>
<td></td>
</tr>
<tr>
<td>neytal</td>
<td>Water lily variety</td>
<td>Coastal tract</td>
<td></td>
<td>Sorrow of lovers due to separation associated with coastal areas</td>
<td></td>
<td></td>
</tr>
<tr>
<td>paalai</td>
<td>Ape Flower</td>
<td>Desert tract</td>
<td>Group of 7 classes of melodies</td>
<td>Temporary separation of lovers associated with desert tract</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Issues in and Implication for KOTS**

The foregoing examples and illustrations, while certainly not exhaustive of the nature of concepts and conceptual relations in Indic culture-specific domains, bring out an important aspect, viz., the unity of ideas cutting across domains. Indic cultures and art forms reveal a “unity, a correspondence of concepts”. For example, *taala* refers to measurement of “rhythm in music, cadence in dance, height in sculpture and area in architecture” (Srinivasan, 2002). Similarly, *ragas* are associated with deities, music, color, gender, mood, and so on. This can be seen across genres of performing arts such as music and dance, across art and architecture. Srinivasan has also brought out the relation between Indian philosophical thought and temple architecture (as seen in Angkor Wat temple in Cambodia). We have also seen the extensive relation between flora and naming of terrains, music, etc. In other words many of these concepts exhibit a trans-disciplinary character. This characteristic could be related to Ranganathan’s proposition of seminal mnemonics implemented in his Colon Classification. Ranganathan’s proposals such as systems and specials also seem to be aimed at accommodating different schools of thought; for example, the idea could be applied when handling different systems of calendars that are extensively used.
An aspect of the nature of these domains that requires attention is the extensive web of relations obtaining between concepts cutting across domains. Reference has already been made to hierarchical relations that are widely recognized and provided for in KOTS. Hierarchical relations as recognized by KOTS are, to a very large extent, based on class membership (class inclusion / generic relation). This in turn is based on the ability to categorize a concept as a member of a group of concepts which share certain common attributes. In other words, the categorization of a concept into a class depends on defining „necessary and sufficient” conditions that every member of that class must satisfy. While this is reasonably straightforward in the domains of science and engineering, in culture-specific domains it is not easy to define such conditions for class membership. As such we end up treating most conceptual relations in these domains as lateral relations. This again leads to imprecision and to inability in precisely expressing the nature of relation between two concepts. An effort, therefore, to categorize lateral relations (non-hierarchical associative relations) into a limited number of categories has been attempted (Neelameghan and Raghavan, 2006). Categorizing lateral relations by itself does not help improve retrieval unless we find mechanisms for implementing these in information systems. A study is being done to explore if these relations could be built into an ontology which in turn could be used in a search interface. For example, the web of relations between the concept kurinji and other concepts is graphically represented in the following Figures:

Fig. 1: Web of Relations
Such a facility should help better navigation among related concepts in a search process. Secondly, it was realized that a complete display of all related concepts is indeed very difficult in these domains given very extensive web of relations obtaining in reality cutting across domains and disciplines. An additional approach that was experimented with was to link the KOT to other lexical tools.

References
Eduardo Ismael Murguia - Universidade Federal Fluminense, Brazil
Rodrigo de Sales - Universidade Federal Fluminense, Brazil

CNPq’s Knowledge Area Table as a Knowledge and Power Apparatus

Abstract
This work is a first reflection on what we understand as knowledge organization based on politics. To do so, we resorted to Foucault’s conceptions of politics, state and governance, aiming to analyze an instrument that guides knowledge organization in Brazil’s research and academic fields. The current version, updated in 1984 by the CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico – National Council for Scientific and Technological Development), of the Knowledge Area Table (KAT) represents and establishes investigation in the scientific and technologic fields. We highlight that the position of Information Science in such Table was a product of national and international reflections in the 1980s on building a scientific subject. Information Science (IS) is placed higher than Information Theory, Library Science and Archival Science in the Table’s hierarchy. This led to a demand to promote interdisciplinarity which, although specific to certain periods, made the discipline acceptable.

Introduction
This work is an initial reflection aimed at expressing our interests as researchers in the fields of epistemology and knowledge organization. To do so, we intend to discuss, and analyze the classification of knowledge adopted by CNPq and the place of Library Science, Archival Science and Information Science in it. Such tables have been used to mark and map the different professions in relation to different kinds of knowledge.

As mentioned by Olsson (2010), only later did Library Science and Information Science (LSIS) researchers adopt Foucault’s ideas. And such ideas were rarely used, if ever, compared to the rest of the social sciences. Since the 1990s, however, Library Science and Information Science researchers have used Foucault as an instrument to critically analyze the institutions and research, and as a conceptual reference to understand informational behavior. Another of Foucault’s features which is crucial to LSIS’s studies is exploring the relation between knowledge and power. Foucault believed power and knowledge should not be seen as separate things, but as products of the same social process.

We must admit that power produces knowledge (and not simply benefits it because it serves power, or applies it because it is useful); power and knowledge are directly implied; there are no power relations without a correlation in the field of knowledge, nor knowledge that does not assume and constitute power relations at the same time (Foucault, 2005 b, p. 27).

What does an answer on what power is refer to? Foucault (2005 a) prefers to divert the question to observe what is at stake, to its mechanisms, its effects, these different power apparatuses and how they are wielded in society’s different fields. Foucault believes power is neither given nor exchanged, nor retaken, but it only exists and is exercised in action; it is a relation of power.

We believe power must be analyzed as something that flows, rather as something that works in chain. It is never here or there, never in someone’s hand, never owned like wealth or a product. Power works. Power operates in a network, and within that network, not only people circulate, they are also in a position in which they can be subject to that power or wield it. They are never a passive target or complier with power. In other words, power flows among people, it is not applied to them (Foucault, 2005 a, p. 35).

The kinds of knowledge that first appeared during the classical period – the 17th century, as mentioned by Foucault (1992), were crucial to observation and classification. The new way to observe established is more important than the desire for knowledge, and that will
be the link between words and speech. The showcase of things, the theatrical way in which objects are displayed, will be replaced by catalogues, the ruling of objects into squares linked by other kinds of relations, i.e., classifications. This is nothing short of a new way of naming what can be seen.

We believe classifying no longer refers to linking objects based on displays, but in organizing its concepts, which will manifest itself in ordering relations designed to equip kinds of knowledge with instruments. In this way, separating the different kinds of knowledge in the Classic period meant the detachment of words and things. Thus, a kind of knowledge is established based on a new way of understanding, the new organizing perspective of modern rationality.

Discipline and Punish (2005b) is one of Foucault’s attempts to codify and materialize (in the sense of physical completeness) the knowledge-power relationships. In this way, he turns to places where power is socially wielded: institutions. In the order established in the 17th and 18th centuries, Foucault elects discipline as society’s elaborating apparatus. Discipline means distribution, control, organization and the composition of power.

Given such premises, we select the university as the institution/place for our investigation. Foucault (2005a) believed science did not exist before the 18th century. There were different types of science, knowledge and philosophy (which was the system that organized and communicated various types of knowledge). Disciplines appeared in the 18th century, embedded in the culture called “science.” In that century, the progress of reason made the University necessary and it took on the role of selecting different types of knowledge. Although universities established a kind of monopoly, as an institutional field, they also had the right to exclude and disqualify knowledge that was not formed within them.

Government Actions for Science and Research: the Creation of Capes and CNPq

Tarapanoff (1993) says Brazilian science was built by sporadic events throughout its history up to the 20th century. Government intervention in the sector led to the creation, in the 19th century, for example, of the Escola de Minas de Ouro Preto, the Instituto Agronômico de Campinas, the Instituto Butantã de São Paulo and to the Instituto Oswaldo Cruz do Rio de Janeiro. These isolated interventions were later replaced by centralized planning. The year of 1951 saw the fruition of the Vargas’s policies which directly met the need for education capable of dealing with the technological innovations needed to improve productivity in several sectors. There was continuity in the policies that established public education, which started in 1930 with the creation of the Education and Health Ministry and the Commission to promote the National Campaign for the Improvement of Higher Education Staff, which led to the creation of the Campaign for the Improvement of Higher Education Staff (CAPES) in 1951. The implementation of the Programa Universitário in 1953 was a result of this. In 1961, CAPES started reporting directly to the president of the republic; however since 1964 CAPES reports to the Ministry of Education and Culture.

Another important development was the creation of the Conselho Nacional de Pesquisa (CNPq- National Council for Research). The CNPq’s goal was to promote and stimulate scientific and technological investigation in any field of knowledge by supporting research and collaboration with other Brazilian or foreign institutions, and organizing and supporting courses to train specialized researchers and technicians. CNPq also proposed to

---

1 Getúlio Vargas governed Brazil between 1930 and 1945 and from 1951 to 1954. The policies mentioned refer to his second administration.
2 Currently National Council for Scientific and Technological Development.
establish relations with national and foreign institutions to exchange technical and scientific documents.

All these developments during the second Getúlio Vargas administration suggest a clear intention of the state to create conditions supportive of scientific research in the form of necessary institutional mechanisms.

I believe the word “governmentality” means a set of institutions, procedures, analyses and reflections, as well as calculation and tactics that allow wielding such specific, although very complex form of power aimed at the people, having political economy as its first type of knowledge, and safety apparatuses as its essential technical instrument. I also understand “governability” as the tendency, the power line in the entire West which has been long conducting the predominance of that kind of power which we can say “rules” over all others. (Foucault, 2009, p. 143-4).

By centralizing funding of educational and research activities, the government also established some control over the kind of research and the disciplines to be supported by the state. Such institutions became the government’s power apparatuses by directing, classifying and planning research and its use nationwide, based on the idea of development.

Classification as a Validation Apparatus within Brazilian Scientific Production

In a revealing article “What is an Apparatus?”, Agamben (2009) explains Foucault’s concept. Foucault himself does not set out its meaning explicitly, although he commented on it briefly:

- I believe an apparatus is a kind of formation whose main purpose was to emerge at a certain moment. The apparatus has a dominant strategic function. The apparatus is always part of a power game.
- What I call an apparatus is larger than episteme. Or episteme, I would say, is an essentially discursive apparatus, unlike the apparatus, that is discursive and non-discursive (Agamben, 2009).

Foucault takes sides in a crucial issue, which is also his own problem: the relations between individuals as living beings and their historic element. By historic element I mean a set of institutions, of the processes of subject constitution, and rules that establish power relations. Foucault’s final goal, however, is neither to reconcile such elements nor to emphasize their conflict. He researches the concrete ways in which statements and apparatuses act on power relations, mechanisms and games (AGAMBEN, 2009).

Foucault repeatedly resorts to the etymological meaning of the word apparatus, highlighting its juridical, technological and military meanings, for which there is always the apparatus indicator, such as rules and orders aimed at causing an effect or confronting a situation that requires practical measures:

An absolutely heterogeneous set that includes discourse, institutions, architectonic facilities, regulatory decisions, laws, administrative measures, scientific statements, philosophical, moral and philanthropic propositions; in short, all that is said and also what isn’t said, such are the apparatus’ elements. The apparatus itself is the network established among these elements (Agamben, 2009 p.29).

In this paper, we propose to investigate and interpret CNPq’s classification table of knowledge areas as a validation apparatus. That does not mean it is the only one, because publications in scientific magazines, summits, curricula, statutes, etc. also are knowledge validation apparatuses. Like any apparatus, classification supports and is supported by the discourses that include it, thus objectifying itself in institutional relations.

We believe the classification tables that arose in the second half of the 19th century, among which we highlight Dewey’s Decimal Classification (DDC), were the first encyclopedic classification proposals based on pragmatic principles. That is, organizing types of knowledge around a decimal notation that can encompass the multiple associations and articulation of the various areas of knowledge. No longer an Aristotelian vision translated into rhetoric hierarchy, but a dynamic and modern effort to organize knowledge.
Unlike Dewey’s intentionality (in the sense of problem solving), Otlet and La Fontaine’s Universal Decimal Classification (UDC) aimed at concrete application. Otlet’s universal catalogue needed a structuring element capable of answering cognitive questions materialized in the combinations and manipulation of files in the universal catalogue, still under construction. Otlet’s new development however, was intended to solve another problem; building the bases of a new knowledge/action which was called Documentation to meet the information demand that science at the end of the 19th century and beginning of the 20th century required.

In short, we can understand the DDC as an a priori hierarchical effort of normalization which could structure encyclopedic knowledge within any bibliographic collection. In other words, an epistemological construction of a validation apparatus of the types of knowledge placed in proposed hierarchies. UDC, on the other hand, consists of an a posteriori effort, which, despite its better theoretical adaptation, emphasizes the validation of a new profession based on the construction of a new object/method.

In Brazil the CNPq was the first institution to establish a classification table called Knowledge Area Table (KAT) in 1976. This was modified in 1982 and 1984; the 1984 version is still in force. This is used by CAPES and other federal and state institutions, with minor changes. Its main goal is management and evaluation of direct public funding of education and research in Brazil. Thus, the difference between the KAT table and the two other tables mentioned above is its intentionality as a political apparatus of government decision-making.

The Knowledge Area Table of CNPq (KAT/CNPq) organizes and classifies the possible knowledge areas in the country in hierarchical levels, in order to manage and assess them. In the last two decades, there were many initiatives to restructure KAT, but the 1984 version is still in effect (SOUZA, 2011).

**Knowledge and Politics: IS’s Position in the KAT/CNPq**

Table 1 shows the divisions of Information Science according to the KAT/CNPq:

<table>
<thead>
<tr>
<th>Classif. Level</th>
<th>1st level</th>
<th>2nd level</th>
<th>3rd level</th>
<th>4th level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broader Area</td>
<td>6.00.00.00-7</td>
<td>6.07.00.00-9</td>
<td>6.07.01.00-5</td>
<td>6.07.01.01-3</td>
</tr>
<tr>
<td>Subarea</td>
<td>Applied Social Sciences</td>
<td>Information Science</td>
<td>Information Theory</td>
<td>General Theory of Information</td>
</tr>
<tr>
<td>Specialization</td>
<td>Library Science</td>
<td>Archival Science</td>
<td>Classification Theory</td>
<td>Quantitative Methods. Library Science</td>
</tr>
<tr>
<td></td>
<td>Archival Organization</td>
<td>Information Retrieval Techniques</td>
<td>Information Dissemination Procedures</td>
<td></td>
</tr>
</tbody>
</table>

Source: CNPq, 2012

Among the nine major areas defined by the KAT, Information Science (IS) is included as one of the 13 Applied Social Sciences sub-areas (Area 6). IS in turn branches out three ways: Information Theory, Library Science and Archival Science. Such specializations are
again divided into more specific areas. When the 1984 table was defined, regarding the IS area and its relations with Library Science, we notice that:

a) In the 1980s, when Lilley and Trice (1989) sketched out a history of IS, they identified two elements in its origin: 1) External forces such as the increase of scientific and technological reports after World War II, and the problems in organization and recovery deriving from these; the continuous increase in funds to promote documentation and organization of information and; some visionaries’ inspired thoughts; and 2) The increase and improvement of users’ services and the emergence of new technologies.

b) At the same time as Meadows (1987), Lilley and Trice proposed that the combination of Bush’s ideas of an artificial memory with Bradford’s ideas about refining and expanding Documentation’s scope, are, if not the origin of IS, at least the transition from Documentation to IS.

c) According to Lilley and Trice (1989), between 1945 and 1985, IS flourished, matured and absorbed documentation, threatening Library Science. Specialists say that IS is ready to join forces with Library Science in a new effort to supply information using the latest innovations in knowledge and technology.

It is clear that Lilley and Trice (1989) believe IS was already a ripe and an established field based on information theory and new technologies. Although Shera (1980) proposed that the term information was taken by the new professionals to differentiate from old Library Science and assert its academic superiority, in the 1980s IS already was a new science, establishing itself as a new discursive formation.

In 1988, Carol Couture, Jacques Ducharme and Jean-Yves Rousseau, wrote the groundbreaking article "L’archivistique a-t-elle trouvé son identité?" on the relationship between IS and Archival Science. This re-examines Archival Science’s traditional principles. The traditional view that Archival Science is an auxiliary science to History, a set of rules and methods aimed to make historical research easier, working as an analysis and critic of historical sources. Modern Archival Science, on the other hand, should focus on managing information, improving and making decision-making easier, increasing knowledge of culture and of organizational processes. Couture, Ducharme and Rousseau (1988) argue that Archival Science must have its place recognized by the society and for this it has to break free from its subordination to other disciplines and must have autonomy to choose alliances with other disciplines.

In the face of this, we would like to stress that, in the 1980s, Archival Science felt the urge to incorporate IS’s discourse, borrowing its objective for the reformulation of the concept of “organic information”.

We intend to set out the epistemological and professional tensions within which the KA/T/CNPq was conceived. Although it was created shortly before the texts presented here, we infer that the debate on IS regarding its relations with related areas was an attempt to question the concrete consequences in the professional and academic areas. In that way, the KA/T aimed to translate the Brazilian academic reflections on the subject to become a political apparatus to “cartographically” ordinate these correlative areas. We also would like to highlight that, in Brazil’s case, such reflections were based on the emphasis given to the studies of the area’s interdisciplinarity, although Saracevic (1996) helped to further the idea when he proposed IS’s appropriations and closer ties with other areas of knowledge. Some authors, however, like Fonseca (2005), believe that:

The interdisciplinarity relations with information science seem to be stronger in Brazil than abroad. The association with information science seems to be an evolution of the archival area in Brazil. It is worth questioning if this is not a “isolated interdisciplinarity,” established as a means to academic survival, given the devastating situation of national archival institutions, but the data shown here indicate that effective interdisciplinarity relations between the areas have been built, considering how similar theses and
Although Fonseca’s (2005) reflection refers exclusively to the field of Archival Science, it reflects the issues we are dealing with and which can be extended to Library Science. Fonseca moves away from her own original idea when she claims that the interdisciplinary relation between Archival Science and IS is due to the fact that graduate programs in IS accept Archival Science’s studies, unlike History’s resistance to doing so. Surprisingly, the thesis that aimed to demonstrate an interdisciplinary dialogue might unintentionally lead to an institutional, i.e. political conclusion.

Understanding the KAT/CNPq as a management and assessment instrument can be further improved if we analyze it also as an apparatus of knowledge validation, which occurs by means of the hierarchical relations established among the areas, sub-areas and specializations. In this way, we can observe that the KAT/CNPq, influenced by the discussions mentioned, sets IS as a sub-area (level 2) of Applied Social Sciences (level 1), which once again indicates its pragmatic nature. For that reason, we also understand, although it may seem awkward, how Information Theory and Library Science and Archival Sciences have equal importance (specialization – level 3). If we reach the fourth level (no classification), however, we can conclude that, due to the number of subjects presented, Library Science showed a far wider range of themes, possibly a reflex of the effective representation of areas when the table was drawn up. We can infer that there are two dynamic and complementary faces between IS’s hegemonic discourse reflected in the KAT/CNPq and the later rules that such reflex will impose on the future of the area’s intellectual and academic production. In other words, what is created as a consequence of a kind of discourse becomes material conditioning.

The table, as a validation apparatus, becomes a pragmatic instrument that guides future IS departments, receiving Library Science and Archival Science in Brazilian universities. We can thus conclude that in Brazil, there is a clear and established regulating and interventionist policy towards research and academic work. KAT/CNPq’s classifying relations establish hierarchy among the different areas of knowledge, leading to the control of scientific production and allocation of resources. In this way, KAT/CNPq consists of a power instrument as it rules, by directing the State’s actions on Brazil’s scientific and technological knowledge. The position of IS, Library Science and Archival Science in the KAT/CNPq, despite its current obsolescence, reflects theoretical tendencies of the 1980s, and still fulfills its role by determining those areas’ scientific production and professionalization.

References
Agamben, G. O. 2009. que é o contemporâneo? e outros ensaios. Chapecó: Argos,
Brasil. Lei 1951. Creates the Conselho Nacional de Pesquisa, among other Actions. Rio de Janeiro, 15 de janeiro de, No. 1: 310


Abstract
The paper presents the approach of using the Functional Requirements for Bibliographic Records (FRBR) model to investigate the complicated sets of relationships among different versions of a classification system for the purposes of specifying provenance of classification data and facilitating collaborative efforts for using and reusing classification data, particularly in a linked data setting. The long-term goal of this research goes beyond the Dewey Decimal Classification that is used as a case. It addresses the questions of if and how the modelling approach and the FRBR-based model itself can be generalized and applied to other classification systems, multilingual and multicultural vocabularies, and even non-KOS resources that share similar characteristics.

Introduction
This paper reports on part of a long-term research project to model classification systems in multilingual/multicultural contexts using general conceptual models to support organization and discovery in Semantic Web settings. The research reported on here focuses on using the Functional Requirements for Bibliographic Records (FRBR) model to investigate the complicated sets of relationships among different versions of a classification system for the purposes of specifying provenance of classification data and facilitating collaborative efforts for using and reusing classification data, particularly in a linked data setting. We limit our consideration in this study to investigation of the whole system and various versions of the system as the focus of modelling; component parts of a version are out of scope for this part of the investigation.

The FRBR model, using the entity-relationship methodology, consists of three groups of entities, attributes of the entities and relationships among them (IFLA, 1998). The three groups of entities can briefly be labelled as products, agents and subjects. Group 1, the set of entities used to model intellectual and artistic products, is the core of the model and consists of the entities work, expression, manifestation and item. The work entity represents intellectual and artistic endeavour at the highest level of abstraction independent of any specific form, language or medium. In our discourse we often identify intellectual creations such as the play “Hamlet” by William Shakespeare as a distinct unit independent of the various translations or adaptations of this work and the work entity generally captures this abstraction. A work entity is in a sense the conceptual “content” that is shared between all different but still comparable variants of this particular play. The expression entity identifies a distinct externalization of a work in a specific language or form. A work such as the play “Hamlet” may exist in different translations and adaptations and each of these can be identified as unique derivation of the same work. Expressions are themselves abstract but an expression is distinguished from work as its nature is the signs that express the concepts of the work it realizes; a German translation of Hamlet is different from a French translation. The next level is the manifestation entity. Manifestations reflect the way content is published in the shape of produced publications. The German translation of “Hamlet” published as a book identified by a specific ISBN is a particular manifestation entity. The same expression can be published in different manifestations and a single
manifestation may contain more than one expression. The item entity is the last entity in the abstraction hierarchy and reflects the actual copies that exist of a particular manifestation, e.g., on the shelves of different libraries. The FRBR model supports four user tasks (find, identify, select, acquire or obtain access).

While the FRBR model is often expressed in the context of bibliographic records, it has also been suggested as a general entity-relationship modelling tool (Zumer, Zeng & Salaba, 2010). A DCMI task group discussed and proposed extending the FRBR modelling to knowledge organization systems (KOS) in building a Dublin Core application profile for KOS resources in a workshop in 2010 (Zeng & Hodge, 2011). Our interest lies in modelling relationships between KOS to support the tasks of specifying provenance of classification data and facilitating collaborative efforts for using and reusing classification data (represented as the four FRBR tasks). We have identified the following questions to guide our exploration of the application of the FRBR model to classification systems:

1. What is the basic “work”? Is it the system as a whole, or an edition of the system? Does this vary by system to be modelled?
2. How should the entities expression, manifestation and item be applied to classification systems?
3. Does the choice of “work” suggest an adjustment to identification methods for individual systems?
4. What relationship does an expression developed at a later point in time from the original work have to the work and to other expressions?
5. How do translations, translations with expansions/contractions, subsets, etc., affect the model?
6. What implications, if any, does the availability of manifestations in structurally different formats have on the relationships among works, expressions, and manifestations?

Using the Dewey Decimal Classification (DDC) system as a case study (in particular, the 22nd edition of the DDC), we explore some of these questions.

Figure 1 contains a top-level view of the different versions of the DDC emanating from DDC 22 (Dewey, 2003).

DDC 22 (Dewey, 2003) was initially published in 2003; the various DDC 22 translations were published in 2005 (German), 2007 (French), 2009 (Italian), and 2011 (Swedish-English mixed version). Abridged Edition 14 (a logical abridgment of DDC 22) was published in 2004; translations followed in 2005 (French), 2006 (Italian and Vietnamese), and 2008 (Hebrew and Spanish). Each translation is based on an updated version of the underlying data file associated with the English-language version of the edition. In the case of nearly every full-edition translation, additional content has been developed for the translation and either added directly to the English-language edition, or represented by a logical abridgment in the English-language edition (Beall, 2003). In several of the abridged-edition translations, content deemed important in the cultural context of the translation has been incorporated from the corresponding full edition. The Swedish-English mixed translation of DDC 22 is a special case: it contains extended Swedish content related to geographic areas and languages; it also contains ingested classes in English imported directly from DDC 22, plus the full English-language index in addition to the Swedish one (Mitchell, Rype & Svanberg, 2011).

---

1 DDC, Dewey, Dewey Decimal Classification and WebDewey are registered trademarks of OCLC Online Computer Library Center, Inc.
In addition to the standard full and abridged DDC editions, there are other abridgments and subsets represented in Figure 1. The top-three levels of the DDC, commonly known as the “DDC Summaries,” have been translated and made available as separate entities in the eleven languages listed in the box on the left-hand side of Figure 1. “Guide” is the 2005 publication *Guide de la classification décimale de Dewey*, a French-language customized abridgment of DDC 22 in outline form accompanied by a instructional text on applying the DDC. *200 Religion Class* (2004) is a subset of DDC 22 containing the full schedules from 200 Religion and 170 Ethics, the Manual notes associated with classes in 200 Religion, and an extended index to 200 Religion. *DDC Sach-Gruppen* is a German translation of selected DDC 22 top-level classes (including some below the three-digit level); its primary use is to organize the national bibliographies of Germany, Austria, and Switzerland (the four languages in the box on the right-hand side of Figure 1 are translations of *DDC Sach-Gruppen* used in the national bibliography of Switzerland).

The view in Figure 1 is format-neutral. For example, the print version of DDC 22 was published in 2003; a web version was made available nearly simultaneously in WebDewey. At the time of initial publication, the underlying database was represented in a proprietary markup language (ESS), and distributed to translators after being transformed into ESS XML (an XML version of the same markup language). The German translation was published in print and web versions in 2005; the German web version, MelvilClass, presents the DDC in a different end user format from that used in WebDewey. In 2009, the DDC Summaries were transformed from ESS XML to RDF/SKOS, and subsequently presented in human (XHTML+RDFa) and machine (RDF) versions in dewey.info, an experimental space for linked DDC data (the machine version of dewey.info has three different RDF serializations: RDF/XML, Turtle, and JSON). In 2010, OCLC moved to a new underlying representation for the DDC, adopting one based on the MARC 21 formats for classification data (to represent...
class records) and authority data (to represent Relative Index and mapped terminologies associated with class records). At the same time, OCLC adopted MARCXML as the distribution and ingest format for DDC data across versions. Also in 2010, a new end-user version of WebDewey was developed that could accept Dewey data in any language represented in MARCXML and display it using a generic user interface. In January 2011, the Swedish-English mixed translation of DDC 22 became the first DDC translation to be published under the new generic WebDewey end-user interface.

In summary, the complicated network of relationships described above illustrates the need for a clear model for describing the relationships among versions. This is important to the whole life cycle of a KOS instance, the management of a KOS family, the mapping efforts that involving a source KOS and maximizing reuse of an intellectual work to support organization and discovery.

We return to our initial set of questions, starting with “what is the basic work?” If we look back at figure 1, “DDC 22” is featured as the central element and thus could be considered the work, surrounded by expressions. The realization of each expression in different published formats (print, web, linked data) are manifestations. The entity item is the print copy of the classification in hand, the accessed web version as it appears on the screen, etc.

The question is, in the current information environment, does it make sense to declare a particular edition as work? While “edition” has been the traditional identifier for DDC data, the situation has always been far more fluid. Editions are updated constantly; translations based on editions usually reflect an updated (and sometimes expanded or contracted) view of the base edition. Should the system itself be the work, and “editions” presented as nested expressions? Is there some other representation? We explore these questions through some user tasks related to linked classification data.

As noted earlier, the FRBR model supports four user tasks (find, identify, select, acquire or obtain access). While it is conceivable that one would want to apply these tasks with respect to versions of the DDC, the more likely user tasks related to linked classification data are focused on the class level (a component part of a version out of scope for the study at hand but a central focus of the FRSAD model).

In dewey.info, the DDC (the system itself) is represented as work at the scheme level; however, most user tasks center on the class level. For example:

Find everything about class 641 in any edition:
http://dewey.info/class/641/

When the result is returned, it is contextualized in terms of the scheme and edition. For example:

http://dewey.info/scheme/e22/ (DDC 22)
http://dewey.info/scheme/a14/ (DDC Abridged Edition 14)

The following are further examples of user tasks associated with linked classification data in a rich environment of classification expressions and manifestations:

Find everything about class 641 in DDC 22 in any language or format:
http://dewey.info/class/641/e22/

Find everything about class 641 in 2009 in any edition, language or format:
http://dewey.info/class/641/2009/

Find everything about class 641 in any edition and obtain it for me in RDF:
http://dewey.info/class/641/about.rdf

Find class 641 in DDC 22 in German and obtain it for me in HTML:
http://dewey.info/class/641/e22/about.de.html
Our tentative conclusion is that it does not matter if the FRBR model is constructed with system or edition as work as long as the relationships among expressions, manifestations, and items are well specified. Classification system designers may find it preferable to move away from a numbered “edition” model to one based on classification system as work with other aspects presented as nested expressions. These aspects could include elements such as type of data [full, abridged, other] and timestamp of data, with language a further identifying aspect.

The DDC case study within the framework of the FRBR conceptual model presents both opportunities and challenges in defining and clarifying KOS relationships. While the DDC is probably the most complicated case to be tested in FRBRization or modelling process, the long-term goal of this research goes beyond the DDC. Questions remaining to be answered include if and how the modelling approach and the FRBR-based model itself can be generalized and applied to other classification systems, other multilingual/multicultural vocabularies (symmetrical and non-symmetrical vocabularies), and even non-KOS resources that share similar characteristics (such as technical manuals). The Linked Data universe provides another dimension for the exploration of the impacts, needs, and use cases in FRBRizing KOS relationships.

References
Dublin, Ohio: OCLC. Dewey.info. Dublin, Ohio: OCLC. Available at: http://dewey.info/
WebDewey. Dublin, Ohio: OCLC. Available at: http://dewey.org/webdewey. (DDC 22 was previously available 2003-2011 in the OCLC Connexion service version of WebDewey.)
Farradane’s Relational Indexing and its Relationship to Hyperlinking in Alzheimer’s Information

Abstract
In an ongoing investigation of the relationship between Jason Farradane’s relational indexing principles and concept combination in Web-based information on Alzheimer’s Disease, the hyperlinks of three consumer health information websites are examined to see how well the linking relationships map to Farradane’s relational operators, as well as to the linking attributes in HTML 5. The links were found to be largely bibliographic in nature, and as such mapped well onto HTML 5. Farradane’s operators were less effective at capturing the individual links; nonetheless, the two dimensions of his relational matrix—association and discrimination—reveal a crucial underlying strategy of the emotionally-charged mediation between complex information and users who are consulting it under severe stress.

Introduction
This paper reports on an ongoing research project which examines the relationship between the principles of relational indexing developed by Jason Farradane and the combinations of concepts that appear in Web-based information on Alzheimer’s disease. An examination of the hyperlinks in consumer information websites shows that while Farradane’s specific relational matrix may prove unsatisfying as a means of codifying hyperlinks, it does provide insight into the relationship between contingent and permanent associations in Alzheimer’s information in particular, and consumer health information in general.

Background and Literature Review
Jason Farradane’s concept of relational indexing emerged formally in 1979, with the first of two articles on the subject in Journal of Information Science. In it, he outlined a matrix of nine operators, plotted along the axes of two principles of concept combination: association and discrimination. Assigning symbols for each, Farradane argued that one could index anything by separating the concepts, and then locating the relationships between them on this matrix of nine operators, ranging from concurrence (a combination of low discrimination and mere awareness of co-existence) to functional dependence (two concepts distinct from each other but associated firmly in a cause-and-effect relationship).

Farradane has fallen so completely off the radar of classification research in recent decades that we can easily overlook his innovative foresight. His contributions to the Classification Research Group have been widely acknowledged, and his scheme for relational indexing continues to command respect for its daring and foresightedness. He anticipated the modern interest in contextualized and socially-situated classification. Like Derek Austin with PRECIS, Farradane anticipated the need for machine-readable systems that could separate the intellectual work of indexing from the mechanical work of assembling the terms and entries. And his attempt to disambiguate complex and ambiguous language constructs for machine processing and interpretation anticipated by at least fifteen years the Semantic Web initiative with its emphasis on machine-understandable information based on unique identifiers that permit inferential logic on defined pieces of data.

Despite these innovations, Farradane has suffered an eclipse in recent years, an eclipse brought about by the failure of his relational indexing system to achieve widespread
adoption. The limitations of Farradane’s matrix of relational operators as an indexing system have overshadowed his contribution to the theory of concept combination. It may be time to revisit Farradane, simply because his fascination with the codifying of relationships between concepts proved prescient. Tim Berners-Lee, the inventor of the World Wide Web, argued similarly for a view of information based not on data per se but on relationships between data elements (Berners-Lee & Fischetti, 2000, p.12). And as the Semantic Web seeks to develop a new standard of Web design which makes these relationships overt and machine-readable, interest is rising in enhancing our ability to encode relationships. The emerging HTML 5 boasts clarified and enhanced linking capabilities (Pilgrim, 2009).

Even as information systems are developing new functionalities for rendering complex relationships in Web design, consumer health information on the Web is growing in size and complexity. Enabling end users to navigate sites which contain massive amounts of information of many different types is as urgent a challenge as ever, and particularly so with information related to Alzheimer’s disease. Alzheimer’s disease is on the rise (Brookmeyer, et al., 1998, 1337), and its potential for overwhelming our health systems in the next few decades is troubling. Information about Alzheimer’s disease is of interest to varied users, ranging from medical and health care professionals to caregivers and those directly afflicted with the disease. It covers a wide range of topics, including medicine, social work, psychology, law, finance and religion. Above all, the causes of Alzheimer’s disease, together with its interactions with other physical conditions, continue to trouble researchers: everyone intimately involved with Alzheimer’s disease is frustrated by how little we know.

In the first stage of this project, which was reported at the ISKO-UK conference in 2011, the author analyzed concept combination embedded in the prose of two Web resources on Alzheimer’s disease: one a professional information guide, the other a collection of diary entries by Alzheimer’s patients. When these combinations were plotted on Farradane’s relational matrix, the professional literature revealed a concentration on functional dependence, reflective of its concern with isolating a cause; it also revealed a tension between association and functional dependence, reflective of the medical establishment’s uncertainty about the relationship of Alzheimer’s with other disorders such as thyroid disease, vitamin deficiency, stroke, anemia and depression (Campbell, 2011, 35). The prose of the Alzheimer’s patients, on the other hand, revealed a focus on the dimensional and action cells. They were comparatively unconcerned about causation; rather, they were very much concerned with centering their lives dimensionally, by relating the stories of their lives in sequence. They were also understandably concerned with action: what influence would this disease have on themselves and their families? Above all, the self-activity operator appeared in the prose of the Alzheimer’s patients: they expressed intransitive verbs, in which they walked, slept, and thought (Campbell, 2011, 38).

These were interesting findings, but they tell only part of the story. As Rebecca Green reminds us, there are three different levels at which one can examine the combination of concepts: at the syntactic level, the level of relationships within the document, and the relationships between documents (2001). Web-based information resources on Alzheimer’s disease are filled with hyperlinks. What is more, proponents of the Semantic Web will argue for the use of new linking functionalities to describe those hyperlinks in a way that is useful for anyone—human or semantic agent—navigating the site for a particular purpose.

This study investigates the concept combinations manifested in the hyperlinks of three websites devoted to Alzheimer’s disease. By comparing these links to both the linking
attributes offered by HTML 5 and the relational operators offered by Farradane, the study poses the following two research questions:

1. Do Farradane’s operators improve on HTML 5 as a means of encoding relationships into hyperlinks?
2. Does Farradane’s distinction between association and discrimination provide a successful means of analyzing the design and linking structures of Alzheimer’s disease websites?

Method
Three websites were selected: the National Institute of Health’s site on Alzheimer’s disease (http://www.nia.nih.gov/alzheimers), the Fisher Institute (http://www.alzinfo.org), and the Alzheimer’s Association (http://www.alz.org). These are the three most frequently cited sources in Medline Plus’s guide to Alzheimer’s disease. All three resources have a detailed and well-documented menu structure, as well as ample links within the textual information, linking the user to other parts of the site. For each site, the menu structure was noted, identifying the formal organization of the site. Then the hyperlinks within the informational text were collected and analyzed, to assess:
   a. the relationship of the anchor text and the target;
   b. a classification of the relationship according to the HTML 5 attributes for the link and href tags;
   c. a classification of the relationship according to Farradane’s matrix.

Site Design and Menu Structure
All three web resources share certain characteristics in their overall design. The very top of the page established the website’s organizational provenance: the Alzheimer’s Association, the Fisher Center for Alzheimer’s Research, and the Alzheimer’s Disease Education and Referral Center (ADEAR). The ADEAR site is placed within a banner establishing it as part of the National Institute on Aging, which in turn is situated within the U.S. department of Health and Human Services.

All three sites have a menu across the top which offers access to the bulk of the site’s information, in a canonical order which most closely resembles a formal classification system. Most of these top menu items link to subtopics, which consist of short prose pieces, many of them containing hyperlinks to other documents.

Beneath this upper menu bar, all three sites have a central section which contains a variety of images, advertisements for publications or upcoming conferences, and portals for donating or signing petitions. In addition, all three sites contain additional menus, most of which link directly to information accessible through the top menu bar, but which are designed to suggest immediate connections between information and a particular need or predicament. The Alzheimer’s Society site, for instance, has a menu with the title “We can help” with the following elements:

- I have Alzheimer’s.
- I am younger than 65.
- I am a caregiver.
- I am a care professional.
- I am a physician.
- I am a researcher.

The Fisher Center contains a series of callouts in its middle section, with phrases including “I have been diagnosed with Alzheimer’s” and “I’m caring for someone with Alzheimer’s.”
Below this center section we find a lower section devoted in all three instances to newsfeeds, as well as a variety of other resources such as FAQ sites, donation buttons or signup services for receiving email news updates.

All three sites, then, offer three ways into the Alzheimer’s information: through a formal menu structure that suggests relationships appropriate to the information; through a less formal structure that suggests a direct line from a question to an answer; and through news and updates that emphasize recent accretions to existing knowledge.

The Links Themselves

The preponderance of the hyperlinks within the textual documents accessed from the upper menu bars on the sites are bibliographic in nature, implying a relative equivalence in terms of subject, but generally a whole-part relationship between the target and the anchor. The ADEAR section called “Alzheimer’s Basics,” for instance, links its first use of the term “mild cognitive impairment” to a later essay that deals more expansively with the term, while another link with the anchor text “read more about what happens to the brain during Alzheimer’s” connects to a separate document in the ADEAR collection entitled Alzheimer’s Disease: Unraveling the Mystery. In general, these links expand on terms and concepts discussed within the anchor document, by linking to three different types of document:

- Topic sheets and online resources, offered within the organization’s website or a parent website;
- Descriptions of relevant print resources available for purchase through the organization; and
- Resource guides containing catalogues and bibliographies of resources for further reading and study.

The links in the center sections tend to provide similar links, and indeed usually link to the same documents available through the top menu bar. However, there is a question/answer quality to many of these links. The link “I have Alzheimer’s” on the Alzheimer’s Association website leads to a page which anticipates a host of responses to that stark statement: a reassurance that the user is not alone, followed by links to resources which imply an answer to the question, “What now?”: advice on self-care, coping with change, getting involved in support groups, and confronting driving and home safety issues.

The newsfeed links typically provide late-breaking news on research findings, either in the area of causation or treatment.

HTML 5

The relationships that prevail on the top menu bars on these websites map reasonably effectively onto the attributes for the `<link>` and `<a>` elements in HTML 5. Although the HTML 5 specification defines a fairly small number of link types, many extensions to the predefined set have been registered, some of them retained from HTML 4, and others developed by numerous Web stakeholders (Microformats 2012). Many of these are bibliographic in nature—appendix, copyright, contents, glossary, help, and bibliography—and could be mapped effectively onto many of the hyperlinks on these three sites.

In addition, HTML 5 specifies explicitly that the `<link>` element can be used for two purposes: for linking to external resources, and for hyperlinking (WHATWG, 2012). A link to an external resource is a link to a resource that is used to augment the current resource; most commonly, the `<link>` element is used to load a stylesheet for viewing the resource.

It is possible that the `<link>` element could be used to specify a particular type of user, along the lines suggested by the list on the front of the Alzheimer’s Society website, to create separate instantiations of the site for someone with Alzheimer’s, for a caregiver, or for a care professional, physician or researcher.
Farradane’s Relational Operators

By contrast, Farradane’s matrix fits the link relationships only to a limited degree. Designed as they were to represent concept combinations within documents, they are less effective in representing links between documents. Many of the links fall into the “appurtenance” category, in the sense of whole-part relationships, and some could be defined as Action relationships, due to their movement from a recognition of Alzheimer’s to a set of concrete steps that can be taken in response to this knowledge. But the operators show the same ambiguity that plagued many who originally tried to use Farradane’s system, and they map uneasily to the relationships implied by most of the hyperlinks on these three sites.

If, however, we move back from the specific operators to the two axes of the matrix—association and differentiation—we find that Farradane’s paradigm shows us other things equally valuable.

The Three Strategies

None of these three sites would win awards for superlative information architecture: dead links appear frequently; the ordering principles across the various menus often show an unwieldy and inconsistent organizational principle; linking practices are inconsistent within the sites. Nonetheless, throughout the linking patterns on these sites, three distinct patterns of linking can be perceived, each allowing for a different strategy of selection.

The first might be called a strategy of “saturation”. In most of the materials on all three sites, the authors anticipate that the reader can take only so much information at a time, and therefore embed more detailed and comprehensive treatments of each topic in a sub-layer of resources, publications and resource guides that the reader can choose to follow or ignore. This practice resembles that of recorded guides at museums and historic sites, in which the surface layer of quick information is seeded with opportunities to go to a deeper, more involved layer when desired.

The second involves selection through the provision of “direct answers”. These links anticipate that the user is seeking a select piece of information to a specific question: Can I still drive? What medications are available? When should I ask my doctor about my memory problems? These links, generally contained in the centre sections of the websites, anticipate a user need, and direct the user directly to the relevant part of the information archive that has been formally classified in the top menu bar.

The third involves a “newsfeed” approach, in which only the latest resources are provided, presumably to top up the awareness of someone who has achieved a comfortable state of knowledge, and wants only the latest additions.

Each of the three strategies enabled by the linking practices on these websites involves giving the user some control over time and sequence. Links can be used to save time (internal links that jump to the relevant part of a large document), to preserve narrative sequence (allowing the user to bypass a link and continue reading a structured explanation), or to extend the time devoted to a subject (allowing the reader to link to a more expansive treatment of the topic).

Even more important, these linking patterns give the user some power to decide how much or how little to learn. And with Alzheimer’s disease, as with other momentous events, this is a significant edge. The combination of these three strategies enables the user to draw a line that separates the “enough” from “too much.”

Farradane makes a distinction between three stages of the associative mechanism: awareness, temporary association and fixed association. This is by no means an abstract principle for those who suffer from Alzheimer’s disease or their families. The movement
from a casual awareness of the disease to a recognition and acceptance of its existence in one’s life is a painful one, and one which individuals negotiate in their different ways.

The line the user draws through selecting which links to follow and which to ignore charts an uneasy progress through the painful middle column of Farradane’s matrix. The association of one’s life with Alzheimer’s is initially a temporary one, one which involves extensive use in the early stages of the dimensional and action relations. It involves establishing the dimensions of the disease: gathering evidence, confirming the suspicion, attaining the diagnosis, establishing a timeline for the progress of the disease. It involves action: arranging finances and legal matters; arranging care; securing the home; embarking on medication; changing the lifestyle.

Conclusions
Farradane’s matrix of relational indexing, then, is an unlikely candidate for the encoding of relational attributes in hyperlinks, and therefore unlikely to make any great difference to web design as envisioned by groups like the World Wide Web Consortium. As a means of analysis of Web practice, however, it has a surprising relevance. When used to analyze patterns of Web linking, it can highlight vexed junctures of information use. In particular, it shows how Web designers anticipate and allow for the user’s uneasy self-association with an unwelcome undesired affliction.

Acknowledgements
The author gratefully acknowledges the influence of the following scholars on his thinking in this area: Timothy Craven, Jacqueline Burkell and Alexandre Fortier.

References
Classifying Web Term Relationships: An Examination of the Search Result Pages of Two Major Search Engines

Abstract

An examination of search result terms (SRT) of two major search engines and the classification of these terms into the three thesaural relationships – equivalence, hierarchical and associative, indicating their occurrence outside of a controlled vocabulary setting and demonstrating a naturally occurring phenomena in language.

Introduction

The management and organization of Web information is a challenge due to its shear size, distributed nature, lack of editorial control and transient qualities (McCown, Nelson, & Van deSompel, 2009). Web search engines make use of two techniques to organize and manage information returned in the search result pages – categories and result page terms (the coined acronym RPT will be used henceforth in this paper). Search engines use categories to organize search results into specified groupings with meaningful headings (Kleinberg 1999). Categories are extensively used in Web search engines and have been found to be valuable in organizing search results (Kules & Shneiderman 2005) and in providing navigational flexibility (Kules & Shneiderman 2008). The figure below displays categories evident in the Google search result page.

RPT is the second technique utilized by Web search engines to organize and manage information returned from user-generated searches. RPT are used to facilitate the processing of results by providing a means of expanding the original search concept. RPT are displayed in the search result pages after the initial search is executed appearing under various headings such as “Related Searches,” “Also Try,” and “Are you looking for” depending on the search engine used. In Google, RPT appear under the heading “Related Searches” while in Yahoo!, they appear directly beneath the search box as the search term is typed. The purpose of this paper is to examine RPT for evidence of classification methods. Evidence of such methods will provide insight into the organizational approaches used by search engines to help manage Web knowledge. The figure below displays RPT found in Google.
Related Studies

Result page terms (RPT) have been the topic of many studies within the Information Retrieval domain. The results of these studies have shown the use of RPT to be an effective means of expanding and exploring initial search topics and assisting the searcher in refining original searches ultimately filtering irrelevant information (Crabtree, Andreae and Gao 2006, and Wu, Madhavan, Halevy 2011). In addition, RPT have been shown to be effective in increasing search result relevance (Bellare et al. 2007, Ma, Yang, King and Lyu 2008, Guo, Xu, Cheng, and Li 2009, and Xue and Croft 2010).

The Study

Four separate single term searches were conducted (April 1, 2011) in two major Web search engines generating a total of one hundred and eleven RPT. The four search terms used were: “hand,” “foot,” “cat,” and “clouds” executed in the Google and Yahoo! search engines. The nature of this study was exploratory and as a result the four search terms and the two Web search engines were selected to serve as a case study of sorts. Only the first layer of the Google and Yahoo! search result pages was examined. The one hundred and eleven RPT were grouped into the three thesaural relationship types described in Dexter Clarke (2001) - equivalence, hierarchical and associative. A description of the three thesaural relationships is provided in the subsequent subsections.

Thesaural Relationships

Equivalence Relationship

Equivalence relationships represent terms whose meanings are similar to the original search concept. For example, both “porcelain” and “bone china” although different in meaning, are related to the concept “ceramics industry” (Dexter Clarke 2001). Dexter Clarke (2001) describes several equivalence types including; common/scientific names, non-proprietary/trade names, standard names/slang, abbreviations/acronyms, lexical variants such as color and colour, inverted entries such as electric cables/cables, electric, terms from different cultures sharing a common language, terms of different linguistic origin, competing terms for emerging concepts or technologies, current/outdated terms, irregular plurals, quasi-synonym, specific concepts subsumed in a broader concept such as flavour/bitterness/sweetness.
Hierarchical Relationship

A pair of terms is assigned a hierarchical relationship when the range of one term wholly includes the range of the other; for example, the terms “lemons” and “limes” are included in “citrus fruits.” The main purpose of a hierarchical relationship is to identify terms that can be used to signify the same concept but different levels of specificity (Dexter Clarke 2001). Hierarchical relationships include “generalization/specialization (hyponym), type/token (instantiation) and whole-part (meronomy) relationships” (Green 2008).

Associative Relationship

When the relationship of terms is not clearly identified as equivalence or hierarchical, it is assumed that an associative relationship exists (Dexter Clarke 2001, Green 2011). Within thesauri, associative relationships are identified with the abbreviation “RT” for “Related Term.” The main focus of associative relationships is to identify additional or alternative terms. These relationships are not based on semantic analysis but rather on the discretion of the creator (Dexter Clarke 2001). According to Green (2011) associative relationships can include lexical similarity, complementarity, facet difference, and world knowledge.

Results and Analysis

The meanings of the RPT were compared to the definitions acquired from the Merriam Webster dictionary (Accessed 4/1/11 from http://www.merriam-webster.com/dictionary/) as well from popular everyday usage. The one hundred and eleven RPT were analyzed and their relationships identified based on these definitions.

Relationships

The results showed that the associative relationship type had the largest number of members, 73 out of 111 RPT. The hierarchical relationship type had 11 members and the equivalence relationship type, 8 members. During the analysis, a fourth grouping emerged. This grouping named “No Relationship Identified (NRI)” had 19 members consisting of RPT that were unrelated in meaning to the original concept. Upon closer examination of this category, two sub-categories emerged. The first sub-category included RPT incorporating part or all of the original search term but containing no relevance to the meaning of the original search term. The second sub-category included RPT that do not contain any part of the original search term and do not share in its meaning. Sixteen out of the total nineteen members of the NRI relationship type were grouped under the former sub-category while only three out of the nineteen members of the NRI relationship type were grouped under the latter sub-category.

Combination Terms

A greater number of terms consisting of two or more words (combination terms) were evident in the Google results than in the Yahoo! results. There were 41 combination terms returned by Google and 32 combination terms returned by Yahoo! However, a closer examination of the number of combination terms to the total number of terms returned for each search engine showed a larger percentage of combination terms in the Yahoo! search results. Out of the total 37 Yahoo! terms returned, 32 where combination terms (86%) and in two of the four searches conducted in Yahoo! 100% of the terms returned were combination terms. In Google, 41 out of the total 74 terms returned were combination terms (55%).

Conclusions and Discussion

Study findings illustrated the grouping of RPT into the three thesaural relationship types defined by Dexter Clarke (2001)—associative, equivalence and hierarchical—as well the
existence of a fourth relationship type that is not identified by Dexter Clarke (2001)—No Relationship Identified (NRI).” Of the four relationships, the associative relationship type contained the largest percentage of RPT. This seems appropriate for the Web domain since RPT are mined from massive database repositories populated by user-generated Web documents and links (Price, Sherman, & Sullivan, 2001, p. 29) and associative relationships are defined (Dexter Clarke 2001 and Green 2011) as those comprising user-generated terms employed in the indexing or retrieval processes.

The NRI grouping contained the second largest number of terms. This group includes RPT that contain part of the original search term, for example when searching for “cat” one NRI term returned was “category,” as well as RPT that are unrelated to the original search term, for example when searching for “clouds” one NRI term returned was “grass.” NRI terms do not possess meaning similar to the original search term. Search results from both search engines showed evidence of NRI terms. Evidence of NRI terms supports past study findings indicating search algorithms often produce irrelevant search term results due to the automatic pattern matching, co-occurrence and near term extracting native to these algorithms (Efthimiadis, 2000, Berners-Lee, Hendler & Lassila, 2001, Bollegala, Matsuo, & Ishizuka, 2007). All equivalence and hierarchical relationships were found within the Google search results. No evidence of equivalence or hierarchical relationships was found when examining the RPT generated from the Yahoo! search. In addition, both search engines showed evidence of combination terms (terms consisting of two or more words such as “hand guns” or “football games”).

Results of the study seem to indicate that although search algorithms continue to produce irrelevant information such as is evident in the NRI grouping, there seems to be a move towards more sophisticated semantically enhanced algorithms that can process complex connections between search and result concepts. Evidence of complex relationships in combination terms and terms grouped under the hierarchical and equivalency relationship types, albeit minimal, gives credence to the conclusion that there seems to be a shift in focus towards implementation of more semantically based search engines. In addition, the presence of the three thesaural relationships in a natural language setting such as the Web seems to indicate that these types of relationships are not limited to controlled vocabularies but are a naturally occurring phenomena in language (Chaffin and Herrmann 1984, Winston, Chaffin and Herrmann 1987, Storey 1993, Morris and Hirst 2004). The current study was exploratory and results cannot be generalized, however results might be representative. A future study of a larger assortment of search engines and the use of single and multiple search terms is needed to further substantiate current study findings. The findings of this study contribute to the field of Knowledge Organization by providing a better understanding of the organization of Web knowledge within search result pages and by contributing to the limited research generated in this area of study.

Acknowledgements
The author wishes to thank Dr. Richard P. Smiraglia, University of Wisconsin-Milwaukee, for his guidance, support and valuable comments and suggestions. The author would also like to thank Melodie Fox, University of Wisconsin-Milwaukee for her insightful comments and suggestions as well the anonymous reviewers for their helpful suggestions.

Work Cited


Rosa San Segundo - Universidad Carlos III of Madrid, Spain  
Daniel Martínez Ávila - Universidad Carlos III of Madrid, Spain  

New Conceptual Structures for the Digital Environment:  
From KOS to the Semantic Interconnection  

Abstract  
Primitive thinking forms affected the organization of knowledge, and at a later date writing also affected organization. Currently, the web requires new forms of learning and knowledge; with the globalization of information, connectivity and virtuality have a bearing on human thought. Digital thinking is shaping our reality and its organizational form. Natural memory, considered to be a process that requires the structure of natural language and human capabilities, is interwoven with a subject and a conscience; memory preserved through writing required other tools to assist it, and classifications, cataloguing, organization or other KOS were created. The new tool for recovering digital memory is the semantic web. This points to information’s future on the Internet and seems to approach the utopia of global, organized information and attempts to give the website greater significance. The Web 3.0 incorporates a proliferation of languages, concepts and tools that are difficult to govern and are created by users. The semantic web seems to be a natural evolution of the participative web in which we find ourselves, and if an effective combination is achieved between the inclusion of semantic content in web pages and the use of artificial intelligence it will be a revolution; semantic codification will be a fact when it is totally automated. Based on this, a collective digital intelligence is being constituted. We find ourselves before intelligent multitudes with broad access to enormous amounts of information. The intelligent multitude emerges when technologies interconnect. In this global interconnection of semantic information an exponential pattern of technological growth can take place.  

Determination of Information Media  
In the postmodern era, the advent of the Internet has brought several revolutions in several different contexts: media, scientific, epistemic and organizational. The new digital materiality transforms science, its contents, and its organizational forms. Information media and writing formats have always affected the way information and knowledge are organized, shaping three fundamental ages in the history of Humanity, the oral period, the written period and the digital period.  

In the oral period, within the oral or tribal societies, totemic societies, pre-literate, and mythic, magic was a form of relation, being societies ruled basically by cosmology and the phenomenology of nature. It is in these societies that the first ideas of social structuring and organization were developed. The social categories in these structures determined the first categories of inclusion and subordination. And, since society is a model in which classificatory thinking works and is developed into multiple aspects, this classificatory thinking was translated into all the categories of knowledge.  

The written age was determined by the appearance of the written alphabet and the development of the polis, which meant a confrontation with the tribe. Writing was one of the fundamental pillars and made civilization possible. The biggest difference between civilization and the most primitive forms of human society then lay on the written language, and writing led to the necessity of organization of all the written knowledge, something also fundamental in the development of humanity as well as the transmission of knowledge and information.  

In a third point, the digital era, with the advent of the digital information media including text, images and sound, the material aspect was also a determining factor in the development of knowledge and information organization. The digital environment increased the necessity of organizing huge amount of information that was generated. Writing, the printing press, and digital information have also brought new ways of
organization. In the same way that the primitive thinking affected knowledge organization, and writing was a determining factor in knowledge organization too, the digital thinking shapes our reality and its organizational forms.

**New Digital Era Conceptual Forms**

Knowledge medium, and the way it is transferred, affects knowledge itself. Digitization is not a mere transition from the techno-culture. Culture, science, and information are superseding their new hegemonic media, in which all the digital entities will be the receivers of the constructed objects, and the digital medium will not only be the tool and the container but also the content itself.

Furthermore, the cyber sapiens or digital natives or post-readers have structured their information in a new and very different way, using a very different kind of thinking that it is constantly supported by a new type of digital memory and a new organizational form attached to the new necessities that they have. We are irreversibly in the third phase of the cognitive schemes, in which they still live with the universes of the writing and the printing press and at the same time we are distancing ourselves from them. New cognitive skills will be spread by the use of digital media, like if they were accessories of a civilization without own memory, transferred and organized in a new digital way.

At this point, a new participative culture appeared in the form of blogs, wikis and social networks, in which the personal creativity operates. There is a new way of thinking in the digital era that embraces the hybridization of materials, formats, and texts. In short, while globalization of information, connectivity, virtuality and hypertextuality have become part of the human thinking, the Internet has required new ways of learning and knowing.

The social web is characterized by the development of technologies oriented toward participation and collaboration among virtual communities. Some of the tools developed have made it possible to establish social networks that comprise communities where users can include their opinions, photographs, and communicate with the other members of their communities (MySpace, Facebook, Tuenti). They can share and download different types of resources such as images (Flick-r), videos (YouTube), books (Google books), as well as participating and collaborating on collective documents (Wikis) and many other types of digital collaboration.

We are experiencing the preeminence of the visual, the digital and the virtual; our environment is visual and this is more and more widespread. Virtual visualization systems, which give the illusion of immersion in the image, are a rapidly evolving field. They are changing our whole reality, which creates new philosophical and ethical problems stemming from the progress of virtual imagery, since it is engendering consequences for our new way of representing and interpreting the world.

The virtual is becoming part of more and more areas. “Virtual reality” refers to computer simulations of the real world through tridimensional images and external components such as a helmet that allow users to interact with the simulation. The idea is to produce an appearance of reality that allows the user to have the sensation of being inside it. Users move through a virtual reality as if they were in the real world. Virtual reality is the most important technological investigation field of the moment. Organizations—whether museums, publishers, media or others—cannot be as they were before their interaction under digital convergence. The mergers between organizations dedicated to the production of books, audiovisual and electronic messages bring with them new habits. Science and information technologies are going to create a great advance in this context - so much so that they are ready to do away with the border that separates the real from the unreal.
From KOS to the Semantic Web

The natural memory, considered a process that requires the structures of the natural language and the human capacities to be developed, is interwoven in the knowing subject and the conscience, for its support, classifications, cataloguing and other KOS have been created, offering an organized representation of conceptual fields and terms of natural language that are the basis for the representation of electronic information.

However, if our memory is starting to be digital to a great extent, its organization is also starting to proceed in a very similar way than the organization of the human memory while digital information systems are trying to emulate the processes of the human mind. The mental process of knowing is not neither linear nor hierarchical, it is more complex, and this means that a new way of electronic organization with associative structures very similar to those of the natural forms is appearing. For this reason, new conceptual structures for the digital environment are needed, and in this vein the web 3.0 is a semantic, associative and relational web which is much more extensive than mere controlled vocabularies, folksonomies and other KOS.

Metadata vocabularies and outlines are public and online (and comprehensible). This means that the elaboration of information and its organization become global, whereas previously their formats and data models made them very exclusive.

Technologies for Semantic Interconnection

The semantic web points at the future of information in the Internet and it seems to be a better approach to the utopia of global information organization. The semantic web is an extensive web, provided of a greater meaning in which the users of the Internet will be able to find answers to their questions in a faster and simpler way than now thanks to a better definition of the information. By providing more meaning to the Internet, and thus providing it with more semantics, it will be possible to find solutions to common problems in the information seeking processes thanks to the use of a common infrastructure, in which it is possible to share, process and transfer information and knowledge in a simple way. The web 3.0 also will bring a proliferation of new languages, concepts and tools that will be developed by the users and might be hard to control. The future semantic web will be the result of the natural evolution of the current social web and attempts of artificial intelligence. Even today, the term web 3.0 seems to be abstract; however, it might mean a huge revolution by an effective combination of inclusion of semantic contents in the websites and artificial intelligence. The semantic codification would be a reality if it could be totally processed using these tools.

Although the semantic web still looks like a far off idea, several new technologies are being developed that allow to give a greater semantic value to the web. These technologies are making digital information more useful and valuable. Among these technologies included are como las plantillas semanticas su implantación puede ser sencilla, the web designer should write semantic templates, gadgets, capable of showing semantic data (templates of presentation) or of collecting user’s data (templates of input) which will be converted into semantic data. Previous findings have shown that with a little training, and without specific knowledge of semantic technologies, graphic designers are able to write fancy web templates in a simple web by using the tools that they are provided with. These templates could be applied to gadgets, small applications that run without the need of an independent application to launch each one, that include different aspects such as weather, time, Wi-Fi signal, calendar, notes, etc. Even those people with limited knowledge of web technologies can use these semantic templates. Another example of this application is the Google gadget that allows insertion of a template on any web page in a simple way. This
gadget can be easily set up, indicating, for instance, the location of the semantic data that are intended to be shown and the template to be used.

Other technologies that make it possible to provide greater semantic value are intelligent agents or the microformats, RDFa and GRDDL. For instance, the microformats were developed to overcome shortcomings of ontologies in the semantic web for the purpose to deliver the information as a greater semantic value. By using the well established standards of the Internet and taking advantage of features of the social web, the ubiquitous web, the mobile web, the Internet of things, the Internet in real-time, multimedia, permanent, 3D… Every website has some semantic information, but what these agents do is try to emulate their interpretation as a human being would do it. Currently, websites are not being developed more easily to read, since this approach is considered impossible, but these agents are designed to be more intelligent.

These intelligent agents, like microformats, are appearing as a response to the challenges of the traditional semantic web. Their objective is to standardize a set of formats in which to store basic knowledge. These microformats might be the best asset in the near future of the semantic web, to help to build a structured information and give some semantic value to it. So far, however, they mean a modest approach compared to the initial pretensions of the semantic web, although they fit perfectly with how the web as it is currently known. If the use and implementation of these technologies succeeds, the future of the web will allow connecting every kind of information, for instance, pictures or bank statements. Current social networks like Facebook and Google+, as we know them, will be displaced by others in which the connections are established between all kind of objects and not only better users, thanks to a better use of the semantic web technologies.

Another technological tool is based on open innovation and linked data; both of them arise within the general framework of the semantic web. The term “open innovation” refers to the open investigation process in a third-party organization. It means giving agents outside the organization itself responsibility for part of the innovation development tasks, sharing data and private information with them. The term “linked data” refers to the method through which it is possible to show, exchange and link data.

But incorporating the environment is only one facet of open innovation. The other is that of allowing someone from outside to develop some element independently. This is known as “inside-out”, which creates benefits from assets that an organization isn’t able to develop. The nature of linked data technologies adapts perfectly to this form of innovation, which is more and more frequent in other fields: allow others to use something that, in reality, an organization does not use or does not know what to do with. Linked data is already a powerful new technological trend that affects various information sectors. It is not only a powerful technological development, but also an innovation component in the information market.

Open innovation can multiply the library services offer. There are many ThinkEpi notes with an engine that is a technological innovation, such as heightened reality or mobile access; or in process innovation, such as the 30300 standards or client-oriented management. Then there is social innovation, such as social networks or crowdsourcing. Linked data is a transition that is necessary for libraries and cultural information projects. Europeana, the digital European library, is paving the way we must follow.

LOD technologies are very suitable for open innovation. Data integration is easier with them. This is not a minor factor in their dissemination, because they reduce barriers. The semantic web’s slogan, “Anyone can say Anything about Any topic” promotes plurality of information, allowing many agents to add information. The open-world assumption makes greater and greater sense, since data cannot be seen as complete or finished—there is
always a possibility of broadening the information or linking it. The future of the network is in greater association. Among some of its primary advantages, teams working interconnectedly would have much greater power to encrypt information; they would make faster searches in gigantic databases possible; they would make possible the development of secure digital products.

The Future of Semantic Interconnection

A collective digital intelligence is being constituted; we find ourselves before multitudes with linked intelligences, with broad access to enormous amounts of information. The intelligent multitude emerges when communication technologies amplify human talents through cooperation. The technologies that are beginning to make the intelligent multitude possible are the messages sent over mobile phones, social networks and the whole body of information transmitted over everyday devices.

We go from interaction to world interconnection; these are connected intelligences, we are in transit toward a web society. The consequences of this amazing world connection can be glimpsed already in the global network of individual intelligences that are linked—one might even ask if this connectedness could create an intelligence much greater than each individual person. This troubling hypothesis contemplates the possibility that right now we are experiencing one of the greatest leaps in our evolution.

In the 21st century, the agent of socioeconomic change par excellence will be technology; this implies a certain technological determinism as a universal law. In this global interconnection of semantic information an exponential pattern of technological growth can emerge, as suggested by Raymond Kurzweil in his provocative and stimulating book *The Singularity is Near*. He affirms that the links between computers will duplicate every year, which will entail tremendous proliferation of technology and a motor for rapid technological change. He suggests the existence of an exponential pattern of technological progress that has been demonstrated throughout human history; this pattern culminates in technological progress unimaginable in the 21st century, which leads to technological Singularity. This is a future event where it is predicted that technological progress will accelerate. Growing semantic interconnection may mean that it is probable that in the future machines will communicate among themselves and replace humans. The name given to this event is an analogy to the gravitational singularity observed in black holes, where there is a point at which the laws of physics cease to be valid, and where convergence on infinite values makes it impossible to define a function.

Technological singularity approaches; it is generated by the exponential, non-linear growth of innovation in information technologies. This exponential growth of computer capabilities may give rise in the future to fantastic new technologies. This process is accompanied by the change that is taking place in all areas of science, as the latter are becoming transformed into digital information.

Once we get to the point we think may arrive, with an intelligence superior to that of humans, a post-human stage would begin. That moment has been named “technological singularity”—it is anticipated that the first artificial intelligence could be built around a computer simulation of a human brain. The singularity is based on the concept of exponential growth and semantic interconnection.

The development of the semantic interconnection seems like something very futuristic; this theory is presented as something very futuristic, but there have been numerous authors who also turned out to be visionaries of what we now experience as everyday. KOS’s were the 20th century’s relational languages, and semantic interconnections will be those of the 21st.
The organization of knowledge has been mediated by oral expression and, later, by writing. Now it will be mediated by digital information and its new semantic interconnection. The medium has shaped the message, and the interconnected digital medium seems to shape not only the organization and structure, but also reality itself.

References
Concept of ‘Time’, Semantic Relationships and Cultural Frames

Abstract
This paper deals with the impact of culture, especially ancient traditions and indigenous cultures, on conceptualization and semantic relationships. Gives examples of representation of the common “universal” concept of „Time,” viewed through select cultural perspectives. Provisions for representing „Time” in knowledge organization tools (KOTs) are examined. Some of the ideas and examples presented arose from our work for developing multi-lingual, multi-culture databases, the related KOTs, user-interfaces, and retrieval processes.

Introduction
Objective
The objective of this paper is to briefly discuss the impact of culture on the formation of concepts about „Time”, interrelationship among concepts and their representation in knowledge organization tools (KOTs) especially dealing with multilingual multi-cultural knowledge resources.

Working Definitions
  
  Concept: A mental representation that stands collectively for all meaningful statements that can be made about an entity.
  
  Culture: Depending on the context „Culture” may signify: (1) a particular society at a particular time and place, e.g. Mayan culture, Vedic culture, Dravidian culture; or (2) the knowledge and values shared by a society e.g. modern Bengali society; or (3) the tastes, styles in art and manners favoured by a community or social group, e.g. Elizabethan poetry, French Renaissance art, and (4) the attitudes and behaviour characteristic of a particular social group or organization, e.g. Jain monks, Ku Klux Clan, South-Pacific Islanders.
  
  Frame: A structure or framework (conceptual or concrete) supporting or containing something.

Background
Concepts, their names (terms), their semantic inter-relationships and organization in meaningful ways are basic to knowledge formation and development of KOTs such as, subject classification schemes, information retrieval thesauri, taxonomies, ontologies, etc. Concepts and their formation are impacted by cultural frames or perspectives. Increasing interactions among peoples of different cultures result in changes in cultural perspectives. KOTs need to accommodate such different cultural perspectives of concepts and relationships among them to enable effective search and retrieval in multilingual domains.

languages of non-Latin origin, such as Sanskrit, Farsi and Tamil. The ideas and examples relating to „Time“ presented in this paper are based on our work on developing multilingual, multi-faith, multi-culture databases, the related KOTS, user-interfaces, and retrieval processes.

**Concept of Time**

The concept of Time is universal although its definition, occurrence, applications (role played) and methods of timekeeping (e.g. calendar) differ across communities and cultures. KOTS such as classification schemes and thesauri may have a separate (common) schedule for Time concepts. In application (e.g. in KOTS) the need to disambiguate the particular cultural way of timekeeping (e.g. in calendars) arises, e.g., 1745 (Hijara), 4345 (Shalivahana). „Time“ may be studied by scholars in different domains; or on the basis of time concept and timekeeping of particular groups, e.g. Sami Time discussed below. Most of the daily, weekly, monthly, annual activities of peoples all over the world are in some manner or the other synchronized with Time. Time is „static“; yet the rotation of the earth and other cosmic phenomena make us feel that Time flows, changes. And there are biological internal clocks to whose functioning – normal and abnormal - the human body is attuned and responds to.

**Time in Indic Cultures**

The notions of Space and Time are extensively dealt with in Indian philosophy and cultures. Indic culture is largely inspired by its ancient philosophy and Vedic culture, even though over the past centuries external invading forces, colonizers, and settlers e.g. Persians, Mughals, Europeans - have had their impact. The different religions and the related practices of Jainism, Buddhism, Hinduism, Judaism, Islam, Christianity, etc., too have impacted. Ancient Hindu units of measures (e.g. timekeeping) are still in use especially in the Hindu and Jain ways of life. Dravidian civilization is the other major Indic culture.

Time (kAla) as conceived and represented in Hindu scriptures ranges from the smallest unit to astronomical units. The Indian perspective of Time is that it is circular or cyclic. The use of the term kAlachakra in Hindu and Buddhist philosophies is also indicative of this perspective of Time. The term chakra (wheel shape) is an important polyvalent sign, organizing metaphor and iconographic device among Indic religions, e.g. Dharmachakra and Sudarshana Chakra. Hindu sages describe the cyclic time as an endless procession of creation, preservation and dissolution. Indian mythology has numerous elements regarding the cyclic nature of time, e.g. the repetitive „Yugas“, the repetitive cycle of the sixty samvatsaras (years), etc. The Hindu cosmology calculates the duration of the Earth as one mahAyuga (about 432 million years) at the end of which the Earth dissolves into the cosmic ocean and remains in that state for an equivalent duration before another cycle of the process of creation begins. The aeons (Yugas) are divided into four periods, namely, kritayuga, trEtayuga, dwAparayuga and kaliyuga. While most cultures base their cosmologies on familiar units such as a few hundred (centuries) or thousand years (millennium), the Hindu concept of time embraces billions and trillions of years. The Puranas describe time units from the infinitesimal truti, lasting 1/1,000,000,000 of a second to a mahamavanvarta of 311 trillion years. There are also several different calendars in use in India, although broadly these are either the solar calendars or the lunar calendars. With a recorded history of over 5000 years, Indian cultures, religions and philosophies have thrown up an extremely rich collection of time-related concepts. The reasons are obvious;
the nature, location and vastness of the country with diversified geography, climate, sea, rivers and mountains, fauna and flora and the exposure to several European and other cultures have, over centuries, served as a fertile platform for nourishing highly philosophic as well as mundane ways of thinking and styles of living; the Online Tamil Lexicon includes over 800 time-related concepts. This is probably true of other linguistic groups also. Narada Samhita (an ancient Hindu text) lists nine methods / types of measuring time and in all the nine systems a year is made up of 12 months, each month having 30 days. Thus, a year is made-up of 360 days. The year is further split into: Ayanas (dakshinAyana and uttarAyana), Ritus (seasons; six seasons are recognized as against four in western calendars), months, pakshas (fortnights), days, mukhrtas, ghatis, kshanas, kalas, kashthas, and nimishas. The concept of Time also manifests in a variety of ways in Indian cultures.

Other Indic Measures of Time: The smallest unit of time is a kaashta which is 18 times the length of time to blink an eyelid. Ten kaashtas make a kshanam ; 12 kshanams make up a muhoortam, and 60 muhoortams make up a day; 30 days constitute a month and 2 months make up a ritu; and 12 months constitute a human year.

In the realm of departed souls - the pitrus - a human month equals the length of a day. The brighter half of a lunar month is the pitru’s day and the darker half constitute their night. In the realm of the Gods or Devas, a human year is equivalent to a single day. The brighter half of the year or uttarayana makes up the day time of the Devas and dakshinayana or the darker half constitutes the night time.

A yuga (epoch) of 1200 Deva years (1200 x 360 = 432000 human years) constitute the kalyugyam or the present epoch; the preceding dwAparayuga was twice the length of the kali yuga i.e. 2400 Deva years; the 3600 Deva years preceding the dwAparayuga made up the trEtAyuga; and 4800 Deva years preceding the trEtayuga made up the krita yuga. A cycle of 4 yugas is called catur yugas. A cycle of catur yugas lasts for 12,000 Deva years or 12,000 x 360 i.e. 4,320,000 human years.

In Indian mythology, these time measurements may be correlated with the process of creation thus:

Brahma is referred to as the creator. A thousand catur yugas make up the day time of a single day of Brahma's life. Another thousand make up the night time of a single day of Brahma. Thus, a single day in Brahma's life spans 2000 x 4,320,000 = 8,640,000,000 human years; 360 such days, each lasting 8.6 billion years constitute a year in Brahma's life, which lasts for a 100 Brahma years. At the end of one Brahma life, another cycle starts. A Brahma's life is also known as a Para. Each half para is a paraardham. It is said that we are currently living in the 2nd half of the life of the present Brahma! In the performance of Vedic rituals, the time-frame in which the ritual is being performed is specified both in macro and in micro terms, the term 'dwiteeya paraardhe' (the second half of Brahma's term) is stated. The reference point here is the moment when Brahma began the creation of the universe. Krishnaswami writes, "When we say 'dwiteeya paraardhe', which Brahma are we referring to? How many Brahmas have preceded the current one? This specification is non-existent in Vedic mantras. Since the whole process is cyclical, with one Brahma commencing when another completes, and with this process repeating forever, there may not be any significance in stating the position of Brahma. If time is postulated as being linear and unidirectional there will have to be an absolute starting point for time. This cyclical nature of time as believed in Indian mythology refers to time as that without a beginning or 'anadi'"

Another measure of time is kalpa. A kalpa or an epoch is made up of 14 manvantaras and each manvantara spans 71 catasyugas. The fourteen manvantaras are respectively swayambhava, savossisha, audhama, thaamaasa, raivatha, sakshusha, vaivasvata, savarni, dukshha savarni, bhramha savarni, dharma savarni, rudra savarni, rouchya and bowdhyya. The present kalyugyam is the 28th in the present Vaivasvata manvantara. Each Brahma's term lasts for 7 kalpas. The current period in time is said to belong to the sweta varaaha kalpa, which is in the second half of the life of Brahma. The puranas are named after kalpas, thus matsya kalpa, koorma kalpa, sweta varaaha kalpa, shiva kalpa, bhringam kalpa, visnua kalpa, etc. Sankalpam: Vedic mantras pin point the time of performance of a ritual - by narrowing down from dviteeya paraardhe (in the 2nd half of the term of Brahma), Sweta varaaha kalpe (in the kalpa sweta varaaha), Vaivasvata manvantare (in the 7th manvantaram), Kaliyuge (in the Kali epoch) - through the finer details such as the name of the current year, month, etc.
More frames of kAlam [time, duration]: kaNam = Shortest duration of time as measured by a snap with the fingers. ilaku = One of 10 varieties of kAlam (time, duration) which consists of 16384 kaNam. Akapatam = One of ten varieties of kAlam (time, duration) which consists of 65536 kaNam, etc. patam = nAzhikai = Indian hour of 24 minutes

New Year: Given the cultural and ethnic diversity of India, New Year's Day is celebrated at different times of the year by different communities; over fifteen variations. Generally the Lunar calendar has been the base of calculations from ancient times. Most of these New Year celebrations are based on the months in the Lunar Hindu Calendar. The New Year celebration may relate to a deity, advent of a season, a certain practice, a cosmological event, etc. Here are a few examples:

- Rongali Bihu (= Bohag Bihu) is celebrated in mid-April. 15 April, (Maanuh Bihu), marks the first day of Hindu Solar calendar. Advent of Spring and seeding time.
- Ugadi in Andhra Pradesh and Karnataka. The first day of the Hindu month Chaitra; advent of spring. (Brahma, the creator, began creation on this day according to Hindu mythology)
- Gudi Padwa is in Maharashtra., the first day of the month Chaitra. Brahma is worshipped and the gudi, Brahma's flag (also called Brahmadhvaj), is hoisted in every house as a symbolic representation of Rama's victory over Ravana.
- Puthandu or Varsha pirappu, in Tamil Nadu., on the first day of the Tamil month Chithirai, 14 April. Women draw patterns called kolams. A lamp called a kuttuvilaku is placed on the center of the kolam, to eradicate darkness. A ritual called kanni takes place. Kanni means 'auspicious sight'. People watch jewellery, fruits, vegetables, flowers, nuts, rice etc., as it is a belief among Tamil people that it brings prosperity. People wear new clothes and special dishes are prepared for the occasion. A car festival is held at Tiruvadamarudur, near Kumbakonam.
- Vishu in Kerala; the first day of the Malayalam month of Medam. Offerings to the divine called Vishukanni arranged on the eve of the festival and consist of rice, linen, cucumber, betel leaves, holy texts, coins and yellow flowers called konna (Cassia fistula). A bell metal lamp-nilavilakku - is placed alongside. People read the Ramayana and visit temples; people wear new clothes, burst crackers, prepare special dishes and the elders of the house give money to children, servants and tenants - Vishukaineetam.

Other New Year days in India include; Cheiraoba (Manipuris); Navreh (Kashmiris); Maha vishuva Sankranti (Odissa); Bestu Varas (Gujaratis); Cheti Chand (Sindhis); Chaitti and Basoa/Bishu (Himachal Pradesh); Pohela Boishakh (West Bengal and Bangladesh); Vaisakh (Sikhs): Chaitra Pratipada (Biharis). Thus,

New Year
RT ugaadi (Andhras, Karnadigas, March)
RT gudi padwa (Maharashtrians, March)
RT puthandu (Tamilians, Chitra, Mid-April)
RT bestu varas (Gujaratis)
RT vishu (Malayalis, Kerala, Medam month)

Similarly for other communities

Each of the days may be associated with a deity or divinity, special offerings, oblations, and other functions and celebrations – giving rise to a network of related concepts.

Similarly, there are differences in the beginning and end of government, business and administrative years, in school / academic year. Venkateswara Rao (2012) writes about the concept of a „spiritual year”.

Time Representation in Space / Architecture
Hindu philosophy, religion and culture which had their origin in India later spread to several countries of Asia in particular. Many elements and aspects of the Indian cosmology, for instance, have been absorbed into the Indonesian Balinese dances and the architecture of Cambodia’s Angkor Wat temples, as in Indian classical dances and temple architecture. The expression of Time in Angkor Wat temple architecture is explained thus: “There were temporary wooden structures built for invoking the Spirit even in the pre-Vedic period,
several centuries before the C.E. ….. The circular fire altar is made up of three hundred and sixty bricks to symbolize a revolution, a concrete representation of cosmic time. The imagery of the cosmic body or Purusha represented by the pole at the centre of the consecrated space-time altar called Mahaavedi corresponds to the human body with all its limbs, senses and apertures. This body in time establishes a relationship with space above and around, thus giving form to abstract speculative thought (Kak, Subbash (2000), quoted in Srinivasan (2002).) We may compare this „unity of ideas across genres” of performing arts, with S.R. Ranganathan’s concept of „seminal mnemonics” and its applications.

Further, in the “Angkor Wat the entire sanctuary including its outer moat is based on Hindu cosmology of aeons or yugas. As mentioned above the duration of the earth is calculated as a Mahayuga lasting 432 million years, which is a single day of Brahma, the Creator. At the end of it, the earth dissolves into the cosmic ocean for an equal period of time, regarded as the night of Brahma. The process of creation, preservation and destruction repeats thus endlessly in a cosmic cycle.”

Sami Concept of Time

A widely studied indigenous culture of Europe is the Sami culture. The concept of Time of these people of the Lapland is based on their strong relationship with nature and the behaviour of the reindeer. It is experiential knowledge “accumulated through repeated experiences of particular situations and arising from the idea of man following the flow of nature without altering it, and that one cannot rush nature.” Their wooden calendar reflects these cultural view points. The seasons are harmonized with the life-cycle - the annual behavioral pattern - of the reindeer marked into eight seasons. In recent years interactions with people from outside Lapland (e.g. trade, tourism, etc) has lead to creating equivalent notations for the Gregorian calendar months in the Sami calendar.

Similar variations in the „calendar” of North American Indians are known. All such cultural variations (frames) relating to the concept of time and timekeeping need to be accommodated in the KOTs. (Neelameghan and Narayana, 2012).

The Sami calendar is also divided into months that more or less match those of the Gregorian calendar, but they are far less rigid about the time period that they define.

Table 1: Gregorian calendar months and Sami months

<table>
<thead>
<tr>
<th>Western</th>
<th>Sami</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>ODDjagemánnu</td>
<td>New Year Month</td>
</tr>
<tr>
<td>February</td>
<td>Guovvamánnu</td>
<td>[Unknown]</td>
</tr>
<tr>
<td>March</td>
<td>Njukčamánnu</td>
<td>Swan Month</td>
</tr>
<tr>
<td>April</td>
<td>Cuogománnu</td>
<td>Snow Crust Month</td>
</tr>
<tr>
<td>May</td>
<td>Miessemánnu</td>
<td>Reindeer Calf Month</td>
</tr>
<tr>
<td>June</td>
<td>Geassemánnu</td>
<td>Summer Month</td>
</tr>
<tr>
<td>July</td>
<td>Siuodnemánnu</td>
<td>Hay Month</td>
</tr>
<tr>
<td>August</td>
<td>Borgemánnu</td>
<td>Molt Month</td>
</tr>
<tr>
<td>September</td>
<td>Čakčamánnu</td>
<td>Fall Month</td>
</tr>
<tr>
<td>October</td>
<td>Golggotmánnu</td>
<td>Rut Month</td>
</tr>
<tr>
<td>November</td>
<td>Skábmnamánnu</td>
<td>Dark-Period Month</td>
</tr>
<tr>
<td>December</td>
<td>Juovlamánnu</td>
<td>Yule Month</td>
</tr>
</tbody>
</table>

As mentioned the Sami concept of Time is based on natural phenomena. The meaning of Sami name Guovvamánnu for February has been lost in time. However, the names of other months show the close connection between Sami culture and Nature. Also, the influence of reindeer herding is evident and the calendar is also based on other yearly occurrences within the reindeer cycle. For example, the Sami name for April, Cuonjománnu, is derived from the Sami word for the hard crust that is beneficial during the migration from the winter grounds to the calving grounds. Thus,
Cuomománnu (April)

SN Term derived from the Sami word for the hard crust, beneficial during the migration of the reindeer from the winter grounds to the calving grounds
RT Hard crust
RT Migration of reindeer
RT Calving ground

The influence of Christianity is evident in Sami name for December. Yule is a pre-Christian word.

The Weeks: The Sami concept of time also includes various names for the different weeks in a year. A few have survived and some of them relate to events of the Christian calendar. The use of calendars designating weeks probably became necessary after introduction of Christianity and weekly church services. The impact of Western society led Sami calendar to place importance on the religious cycle as well as nature’s cycle, a step towards the assimilation of the Sami culture. Still, dependence on reindeer husbandry demanded accurate knowledge of the seasons and life-cycles of the reindeer. As tax collectors, traders and missionaries moved into Sami territory, these weekly calendars became a necessity as interaction with the outsiders needed a more fixed timetable.

The Sami Wooden Calendar: The wooden calendar is also divided into weeks and serves as a portable device for telling time. Fabricated from wood or reindeer antler, this calendar was used to keep track of both natural phenomena and religious occurrences. Written in the runic alphabet, the wooden calendar was a useful tool in helping preserve the balance between nature and religion. These calendars date back to the mid-1800’s and are another indication of the impact of Christianity upon the Sami. In the drawings of the Sami wooden calendar, there is a great deal of detail and attention given to the intricacies of this way of telling time.

The following list of Sami names of the weeks and the wooden calendar are indicators of this change in the Sami culture. In Table 2 below the Sami name (first column) has an Associative Relationship (RT) with the corresponding name of a Saint or Christian event, or Season or Animal etc. (second column).

<table>
<thead>
<tr>
<th>Sami Name</th>
<th>Meaning (Week)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bar‘debei-våk’ko</td>
<td>St. Bartholemew (Aug. 24th)</td>
</tr>
<tr>
<td>Birgit-våk’ko</td>
<td>St. Birgit’s</td>
</tr>
<tr>
<td>Dal’ve-våk’ko</td>
<td>Winter</td>
</tr>
<tr>
<td>Giil’d-Marija-våk’ko</td>
<td>Virgin Mary</td>
</tr>
<tr>
<td>Hallemes-våk’ko</td>
<td>All-Saints</td>
</tr>
<tr>
<td>Mattusmes-våk’ko</td>
<td>St. Matthew’s</td>
</tr>
<tr>
<td>Miccamar-våk’ko</td>
<td>Mid-Summer</td>
</tr>
<tr>
<td>Mikkalmes-våk’ko</td>
<td>St. Michael’s</td>
</tr>
<tr>
<td>Oar’re-våk’ko</td>
<td>Squirrel</td>
</tr>
<tr>
<td>Simmun-våk’ko</td>
<td>St. Simon’s</td>
</tr>
<tr>
<td>Sobmer-våk’ko</td>
<td>Before Leaf-fall</td>
</tr>
<tr>
<td>Gássa-våk’ko</td>
<td>Goose</td>
</tr>
<tr>
<td>Urba-våk’ko</td>
<td>Trees Leafing</td>
</tr>
</tbody>
</table>

Symbols: Crosses denote days of religious significance while fish and leaf sprigs denote various events in nature. By keeping track of time in this way, the Sami can easily refer back to earlier times of the year in order to predict when the fishing season would be most bountiful or whether or not spring would arrive early or not. This sensitivity towards the ways of the natural world was and is mandatory for the reindeer herders, fisherman and farmers alike. Here again each symbol (and name of symbol and picture) can be related to corresponding events in nature.
Concluding Remarks

Hajime Nakamura (1993), writes that “the concept of Time in Indian thought is quite different from that of the West. “In Indian thought, time, like other phenomena, is conceived statically rather than dynamically…. although it is recognized that the things of this world are always moving and changing. But the substance of things is seen as basically unchanging, it’s underlying reality unaffected by the ceaseless flux. The Indian does not concede that we never step into the same river twice; he directs our attention not to the flow of water but to the river itself, the unchanging universal. Indian thought places a high value on universality, and the connection between this, and the static conception of phenomena, is not accidental. "The one remains, the many change and flee. ... this static conception of time permeates Indian thought... it is present in the very forms of language itself, conditioning all philosophical thinking. Nakamura provides examples from Sanskrit language."

From the brief descriptions in the examples given above, we may note:

- Lateral relationships of Time to concepts in different domains – Anthropology, Cultural studies, Sociology, Architecture, Temples, Philosophy, Cosmology, Mythology etc.
- In KOTs, such as classification schemes a common Time schedule is given (e.g. DDC, LC, CC).
- Wherever Time concept occurs its code / term may be picked up from this schedule and attached in constructing the subject class code.
- Classification schemes enumerate Time element in the schedules for a broader domain. For example in Colon Classification Time is enumerated in the schedules for Physics (in the Personality facet) along with other Fundamentals–Space, Energy, Matter, etc; in Cosmic hypothesis, Mysticism; Time in Indian Classical music or Time based theory of music where Time occurs as an attribute (property) of music; Time reckoning in Mathematics; Mental chronometry in Psychology; Time table in Education
- Time may occur as a qualifier (Speciator), e.g. World War II-1939-1945; Work-Time; Birth-Date
- Time qualified (speciated) by another concept, e.g. Time-Auspicious; Time-Wasted
- There are documents that comprehensively treat Time, e.g. Time and the sciences …. / Josefina Mena Abraham and Frank Greenaway, eds. Paris Unesco; 1979; Natural philosophy of time / G.J.Whitrow; 1980; Time in history: views of time from pre-historic to the present day; 1989; The geography of time: the temporal miadventures of a social psychologist, or how every culture keeps time just a bit differently / Robert Levine; 1997. CC provides for such comprehensive treatment in Generalia Entity Study.

Issues to be considered in designing multilingual KOT:

1. Identifying equivalent / near-equivalent concepts/terms in all the languages of the KOT. If such multi-lingual (online) dictionaries are available or prepared, the system may be enabled to pick up all the equivalents (and near-equivalents) from the dictionary for matching with the search term(s). In some cases exact equivalent term (phrase) may not be available in a particular languages, but lengthy definition or scope note may be available. If the terms are coded, say, the class code in a classification scheme, then it can be assigned (used as a tag) in the exact terms, near-equivalent terms or definitions in all the languages. Such a code used in the search expression can pick up the terms, definitions, etc., irrespective of the language of the text. In earlier papers it was shown that in a multi-lingual thesaurus, using a multilingual dictionary (English-Tamil-Kannada), with the search term can be in any one of these languages, the system will pick up equivalent term(s) in the other languages, search and retrieve matching records in all the three languages (Neelameghan and Lalitha, 2011). A similar facility is available in the cc Kannada-English version. In a pilot project, generating a thesaurus of the concepts/terms in the Personality facet for the subject „Education” in the cc Kannada-English version, with facility for searching by the class code (class number) of the search term retrieves all records matching with the code, irrespective of the language of the record / text. (Sharada and Lalitha, 2012, under publication). The code may be from DDC or any other appropriate classification or coding scheme.
2. As pointed out above, the lateral (RT) links to a concept may be different in different cultural frames.
3. Disambiguation of homonyms and homographs.
4. Hyper-linking of terms in a KOT, e.g. Classification scheme, thesaurus, to other KOTS, data collections, images, etc., must be provided to enable comprehensive search and knowledge discovery.

References
Laurence Favier - University of Lille 3 - Laboratoire Geriico, France
Widad Mustafa El Hadi - University of Lille 3- Laboratoire Geriico, France

From Text to Image:
The Concept of Universality in the Knowledge Organization System
Designed by Paul Otlet and the International Institute of Bibliography

Abstract
This paper will examine the documentary universalism defended by Paul Otlet, as an aspect of his universalist philosophy. We will particularly be focusing on two elements of his knowledge organization theory: text and image documents. Then we will compare Otlet’s universalist approach with the challenges of the “universal search” at the internet age. In order to discuss Paul Otlet’s legacy, related to universal search and UDC role, we are currently analyzing the UIR archives, and more broadly, the Mundaneum’s archives. The most important part of our research carried in the Mundaneum’s archives will concern the Universal Iconographic Repertory (UIR) and what it reveals about the nature of the UDC and its’s role in the universal search. This term covers both search engines and online libraries’ catalogs tailored for searching on the Web.

Background
The definition of the modern notions of document and information by Paul Otlet in the XIXth century and the creation of his pioneering Universal Bibliographic Repertory (URB), seem to be associated only to the written document. As Rayward (2003) noted it:

“For him, knowledge was embedded in documents which objectified it and gave it a kind of public status. But documents not only consisted of the written or printed word. Objects, pictures and illustrations, music-whatever had evidentiary value, that “documented” something, were documents. The burning question was how could documents in their many manifestations and formats be made not yield up the knowledge that they “contained” or represented?”

This paper will examine this “burning question” related to the documentary universalism defended by Paul Otlet, as an aspect of his universalist philosophy. We will especially be focusing on two elements: text and image documents. Then we’ll compare Otlet’s universalist approach with the challenges of “universal search” in the internet age, universal search that both search engines and online libraries’ catalogs intend to reach on the Web.

While The International Institute of Bibliography (IIB) was developing the URB, it undertook at the same time a research on iconographic document and ended up creating the “Universal Iconographic Repertory” (UIR). This Repertory contained a wide variety of elements: illustrations1, pictures (photography was just beginning to be developed at the time) but also schemes 2. The main objective of the UIR was to give an idea of the “entirety of Human activities” that could be represented by images, illustrations, graphics, etc. The two Repertories, the bibliographic one and the Iconographic one were not merely juxtaposed. They were linked by a common cataloguing framework and one and the same indexing language, the Universal Decimal Classification (UDC). We'll highlight that UDC’s role associated with the “monographic principle” could not only support a system close to the modern hypertext, as Rayward (1994) showed it, but also served as a pivotal language allowing for what we call today “interoperability” between documents types.

1 See Eugen Wüster’s (Vienna School for Terminology), first work on an illustrated multilingual dictionary published more than half a century after Otlet’s UDC. The Machine Tool, London 1968. An interlingual dictionary of basic concepts; comprising an alphabetical dictionary and a classified vocabulary with definitions and illustrations. Prepared under the auspices of the United Nations Economic Commission for Europe and under the direction of Eugen Wuster, probably inspired by Otlet’s ideas.

2 The first lithography by Nicephore Niepce was done in 1812.
In order to discuss Paul Otlet’s legacy, related to universal search and UDC role, we are currently analyzing the UIR archives, and more broadly, the Mundaneum’s archives.

I. The Documentary Image and the Universal Iconographic Repertory

I-Images as a Source of Information

The Universal Iconographic Repertory’s aim is to systematically record, on a detailed classification basis, the totality of images representing human activities. It is a kind of image encyclopedia necessary for completing illustrated publications, simple demonstrations and descriptive textual elements found in texts, books and periodicals used in teaching and conferences (IIB 1906).

The UIR has been managed by the International Institute for Photography, together with other collections related to photography, catalogues, industrial and commercial leaflets as well. This approach prioritized photos since they were initially an authentic and impersonal reproduction of real things/objects (op.cit.). Since then, photographic manipulation and the later creation of synthetic images has considerably relativized that “objectivity”. Unlike the critical tradition concerning images which goes back to Plato, the IIB members took into account images modeled on the new photographic technique which are supposed to carry a truth value that a written text could not be relied upon to express:

“texts are only opinions while a photo is the literal transcription of fact” (op.cit.).

Selecting iconographic documents to be added to the UIR is based upon one criterion: the objective dimension of the image. The Repertory’s scope will thus cover

“not only photos but also figures and sketches of various kinds (engravings, lithographies and so on,) that could have a documentary interest, and this, whether the image is reproduced by some technical mean or whether it exists only as an original” (op.cit.).

Later (1931) Otlet developed a more elaborate theory of images in which he made a distinction not only between text and image but also between abstract and representative images. He also created a distinction between schematic images and real ones. According Estivals (2010), he was probably the first to recognize schematic images as a different category from figurative ones.

It is quite important to insist on the dual nature of images in the IIB project: first it functions as a book illustration; and then it is also considered as a specific source of information not to be mixed with the book it illustrates; it is also a document which is supposed to be analyzed in itself and for itself.

I-2 The UIR Objectives

The UIR is not conceived as a process only designed for the updating of knowledge but also as a tool for whoever would like to have an idea about a subject, whether he be a researcher or an industrialist. It can also help in preparing a trip or discovering new ground. The first purpose is to keep images in order to help future historians. The witnessing value of an image is essential.

As we’ve already noted, images are also abstract representations, visualizations of ideas. Smiraglia & al (2011) and Acker (2010) analyzed these visual representations. In short, they are not only the information contained in a primary document (photography, postal cards) but they are also elaborated information.

To illustrate the role of these repertories inside his whole project Otlet imagines a pyramid (figure 2) named “Documentation Organization”. The repertories (including the UIR) would be the second layer, the first one an international library, the third a documentary encyclopedia and the last one, the top of the pyramid, is named synthesis. To each layer corresponds a kind of documentary information: primary document for the library layer, bibliographic cards for the second one, scientific and technical files for the
encyclopedic layer, and information representations and visualizations (abstract and tables) for the last layer. Acker (2010) presented Otlet’s encyclopedic project as “a graphic scenographic construction” designed for ideas demonstration and not only for knowledge compiling.

So image has a real importance in Otlet’s universalism theory. The link between knowledge unity and documents diversity, including images, was provided by the Universal Decimal Classification (UDC).

**Fig. 1: Photogrophy of the archive “Document Organization” according to Paul Otlet (©Mons Mundaneum Archives)**

II. The Oneness of Knowledge and the Interoperability between Documentary Repertories (RBU and RIU): the Universal Decimal Classification’s Role

Rayward (1994) stated that:

“Otlet and his colleagues were also responsible for the development of what we would call a highly flexible data management system for databases created from cards and sheets. This was the Universal Decimal Classification, the first of the great modern faceted classification systems which grew out of Otlet’s discovery in 1895 of Melvil Dewey’s Decimal Classification and his recognition of its potential for mapping knowledge domains and encoding complex subject descriptions”.

We are analyzing here how this “data management system” could have been applied to images through Universal Iconographic Repertory and what it reveals about the UDC’s role.

II-1. Classifying Images according to the “Monographic Principle” and the UDC

In the same way as bibliographic documentation is delineated, images were also catalogued on individual normalized cards. Illustrations had to be displayed separately from their context and only the title and the index had to mention the original context. Displaying a continuous series of illustrations on the same card had to be avoided, and each card had to carry one single piece of information.

The printed label at the back of each card had to include three information fields:

- Title
- Identification i.e. entry number, source, date of the document
- Classification: subject matter, persons, place, date

Classification indexes had to be organized according to UDC rules. Iconographic document indexing had to extract all the relevant subjects, selecting them according to the possible future uses of the document. The anticipation of the users’ needs constitute the main indexing principle for images. Indexing has to link users’ needs and knowledge universe expressed by the UDC.
Among these relevant subjects, the “dominant” one, i.e. “the one representing the most interest”, is emphasized in order to file the original document and make it stand out from the rest.

There was a mention saying that:

“doubles will be used for classification on a different basis from that of the classification of the first picture” (IIB 1906).

This is one of the rules allowing to circumvent the strict synthetic and hierarchic nature of the classification used for indexing images. The multiplicity of the topics tied with clues justifying the use of documents doubles in the repertory and the creation of the relation sign (the double point “;”) which didn’t exist in the Dewey system, are ways to overcome the mutual exclusion principle of knowledge organization categories.

Established in this manner, the Universal Iconographic Repertory represented “an ideal collection prototype” (op.cit.) according to the IIB. But the generalization of these indexing rules never took place. Neither museums nor libraries accepted to adopt this model on a large scale.

II-2. Universality of Knowledge and Interoperability of Repertories

According to its authors, the UDC’s purpose was not, to substitute itself for others existing classifications. The debate between forms of specific and universal classification already existed. It is worth noting that Joseph Vallot, who had created a separate system of general iconographic classification for the Documentary Photographic Association of the Paris Museums, ended up by adopting the CDU.

According to the IIB (IIB 1910, Publication n°114),

“The decimal classification system allows a classification on the basis of topic (…). Adopting it does not aim at replacing the special and topical forms of classification which can still be used if they offer a specific advantage. It should rather be used as a system of correlation between all the existing classification systems by offering an international and universal system through which all existing system can be connected. It would be sufficient to establish correlation tables”.

The CDU’s universality was not thought of as the only possibility existing but rather as a form of pivotal language enabling the translatability of all the others. This view brings us closer to the present attempts at creating interoperability between the formats of the metadata through the semantic Web project (RDF format). Van den Heuvel (2009) already noted that:

“Whereas Otlet’s knowledge construction is built up by decimal classification numbers, the Semantic Web uses the Resource Description Framework (RDF) to present and exchange data in a standardized way. Similar to the RDF data model, the Universal Decimal Classification makes statements about the resources. The UDC does not just order subjects or topics in classes by numeric codes, it also allows for linking to additional facets, such as place, language, physical characteristics via its auxiliary tables of connector signs (Rayward 1994). This resembles the linkage in so-called RDF triples”.

This idea is also revealed through Smiraglia et al (2011) discussion of Beghtol C. ideas. Beghtol drew a contrast between the Dewey Decimal Classification (DDC) and Ranganathan’s Colon Classification (CC) in order to describe a transition in 20th century classification theory from top-down “universe of knowledge” systems towards bottom-up “universe of concepts” systems. According Smiraglia et al:

“The “universe of knowledge” approach tends to view knowledge as a whole that can be partitioned into hierarchical structures, whereas the “universe of concepts” approach tends to view knowledge as consisting of conceptual units that can be combined by means of relationships into more complex wholes. Both, however, posit that there are elementary structures of knowledge that can be expressed in a KOS”.

The authors show that “However, her review of the transition from a “universe of knowledge” to

“universes of concepts” did not take into account the Universal Decimal Classification (UDC). There is, however, good reason for according greater attention to the UDC within such a narrative, for it combines, in a singularly perspicacious manner, the enumerative “universe of knowledge” and faceted “universe of concepts’ perspectives”.

They explored three aspects of elementary structures of knowledge that are critical for mediating between “universe of knowledge” and “universe of concept” KOSs, taking the UDC as their point of departure. The third one is

“Interactions between elementary structures of knowledge — that is to say, the interaction between the UDC and other universes of knowledge and of concepts” (op. cit.).

This investigation is essential to our assumption: the very structure of the UDC makes interoperability between different KOSs possible. This work is comparable to what has been done by Heuvel, & Salah (2011 p. 283). In this article the authors reported for experiments undertaken in the Knowledge Space Lab in which main categories of Wikipedia were mapped on the UDC. Although these experiments tried to reveal the faceted structure’s flexibility comparing UDC and Colon classification. This reveals, moreover, the capacity of the UDC to be a pivotal language enabling interoperability.

III-The Documentary Universalism: Paul Otlet’s Legacy

In this section, we focus on Paul Otlet’s legacy related to documentary universalism in the web environment. Two aspects are examined: i) multidimensional features of UDC and web information retrieval requirements and ii) universal search’s claim of search engines at the internet age where all information retrieval systems (libraries’ catalogs, special databases, search engines indexing open web) are interconnected.

III-1 Multidimensional features of UDC and Web Information Retrieval Requirements

The very nature and principles of the UDC i.e. its expressivity through a large range of notational symbols are conducive to searching in automated systems as Mcllwaine noticed (2007:5). Furthermore, the UDC is considered as flexible and adaptable to many applications and situations:

“The attraction of the systematic approach, notated by symbols, rather than an alphabetical approach that is based on one language are considerable for the international retrieval and exchange of information”, (for a detailed account on the UDC nature, see Mcllwaine 2007).

This is one of the major goals of the UDC: its capacity to be a universal notational system bearing in its very nature the possibility of multilingual and multidimensional indexing and retrieval capabilities.

Many authors examined the UDC multidimensionality and how P. Otlet handled it in his universal classification. Heuvel, in an earlier article (2008) quoted by Smiraglia, Heuvel & Dousa (2011) made an interpretation of one of Otlet’s drawings which we think carries the idea of multidimensionality. In this article the authors argued that in Otlet’s mind “the UDC played a central role in a process that combined the top down movement of universe of knowledge systems and the bottom-up movement of universe of concept systems described by Wilson”, (2011: 27). They described Otlet’s drawings in the following:

“(…) these drawings are just a few of many visualizations in which Otlet represented the relationships between elements and the whole of documents and of documentary institutions”.

Otlet’s drawings lead to an interpretation by the three authors that goes in the same direction of our assumptions about the inherent multidimensional nature of the UDC:

“(…) These visualizations reveal how Otlet tried over and over again to create interfaces that could handle the complexity of the multidimensionality of his classification system and so come to a meaningful integration or synthesis of knowledge in his KOS”, Heuvel, Smiraglia & Dousa (2011: 29-30)

The multidimensional nature of the UDC has been compared by Heuvel & Salah to the synthetico-analytical nature of Ranganathan’s faceted classification. Heuvel & Salah (2011:...
283). They mentioned an unpublished manuscript for Otlet “théorie sémantique de la classification” 1908. This manuscript includes sketches that demonstrate “an exploration by Paul Otlet of the multidimensional characteristics of the UDC”. The faceted nature of the classification is an important asset of this classification in the new digital environment. This asset has been also reported for by Schanhorst, Salah, Smiriglia (2011).

The multidimensional nature of the UDC combined with its language-independent nature and its capacity to allow for interoperability could be a valuable asset for its exploitation in the open-linked data era.

This idea is also revealed through Smiraglia, Heuvel & Dousa research work (2011: 26) where they considered the UDC interaction between elementary structures of knowledge (see also Beghtol, 2008-43, quoted by the authors). They suggested the following idea:

“The “universe of knowledge approach tends to view knowledge as a whole that can be positioned into hierarchical structures, whereas the “universe of concepts” approach tends to view knowledge as consisting of conceptual units that can be combined by means of relationships between into more complex wholes” (2011: 26).

The authors explored three aspects of the elementary structures of knowledge that are critical for mediation between the universe of knowledge and “universe of concepts” KOS systems taking the UDC as a point of departure”.

The interaction between elementary structures of knowledge, or the interaction between the UDC and other universes of Knowledge and concepts discussed by the authors (2011: 26) is essential for our assumption: the very structure of the UDC makes interoperability between different KOS systems possible. This work is comparable to what Heuvel & Salah wrote (2011: 283). In their article the authors reported for experiments undertaken in the Knowledge Space Lab in which main categories of Wikipedia were mapped on the UDC. Although these experiments tried to reveal the faceted structure’s flexibility in comparison between UDC and Colon classification this reveals, moreover, the capacity of the UDC to be a pivotal language enabling.

III.2 A century after Paul Otlet, Universal Search at the Internet Age

Universal search became, a few years ago, a commercial slogan in the search engines’ industry. Since Google announced in 1997 that the most important change in “searchology” (the word is by Google) would lie in universal search solution (USS), the information retrieval’s challenge became to transform “vertical search” (i.e. separate search for each type of document: images, videos, news, patents etc.) into a single search box solution, no matter the format of documents.

In libraries, federated search engines were supposed to reveal to the users all the resources the library offers to them (databases, e-journals, e-books) as well as content available on the open web. Today, Web scale discovery hold the promise to fundamentally improve federated search, overcoming its main disadvantages: real time search emphasizing only the most recent documents and, more generally and no particular order of relevance (De Groote & Appelt 2007, Woods 2010). By preharvesting and centrally indexing content sourced across multiple silos, Web scale discovery streamline end user discovery and delivery of content.

This single search box interface to pre-indexed documents, coming from various sources converging on the Web, can be compared to the Paul Otlet project of the Mundaneum. Mundaneum was defined by Otlet (1935) as “an Idea, an Institution, a Method, a Material Body of works and collections, an Edifice, a Network”: the unity of the project reflects the unity of the world; its product is universalism. Today the Internet is for us all theses aspects together if we accept that material cyberinfrastructures replaced the Edifice. Here is (figure 2) a representation of the Mundaneum: at the top, you can read “an Idea, an Institution, an


Edifice, a Network, a Method- For representation and visualization of the world (…)”; at the bottom, you can identify a representation of the Network (Rete Mundaneum). On the left side, the institutions organize and feed the system; on the right side, documentation realizes it.

Based on this Otlet’s representation of Mundaneum, we can say that the main challenge to perform universalism today in the Web retrieval system is the method.

The method aims at integrating both every document type and all virtual spaces converging to the Web. Therefore library catalogs can’t be anymore a stand-alone database neither the Web can’t be a specific document repository. Bibliographic records have to be a highly hyperlinked data set interacting with information resources on the Web. This approach has shifted the emphasis from document to data. After being focused only on metadata formats and ontologies (a kind of domain representation), the method is now “data centered”. Open linked data represents the ideal to make interoperable different subject vocabularies (at least, the major classifications and thesauri) for cross linking between heterogeneous collections. For some authors (eg. van Hooland & al. 2011), “Re-using these established terms for indexing cultural heritage resources represents a big potential for Libraries, Archives and Museums”.

Following a similar approach, some professionals (Granados Collillas 2011) has proposed an indexing system that uses LCSH (Library of Congress Subject Headings) as descriptors, breaking the precoordinated strings and the UDC as a universal ontology.

The project to link subject vocabularies is not very far from Otlet’s approach of the UDC role (see infra chapter II-2) as a “pivotal language”. Twenty-four years ago, in 1988, Eric de Grolier wrote⁴ (de Grolier 1988) that we were still a long way from reaching Otlet’s goals for unifying subject vocabularies. Time has changed and some Otlet’s dreams related to knowledge universalism become realistic challenges.

**Fig. 2: Mundaneum and Cité Mondiale representation.**

Concentration and Research Perspectives

Recently conducted research on Paul Otlet’s 19th century knowledge organization theory shows how his work is relevant and can be a basis for the most advanced development of today digital society. The most important issue is universalism examined from many respects: philosophical, political and technical calls for a comparison with what we call today “globalization”/modialisation, a term used by Otlet. Concepts such as

---

⁴ “(…) à six ans du centenaire de cette date de 1896, qui, avec le rejet de la CDU pour l’International Catalog of Scientific Literature (ICSL) marqua le premier échec du projet otléien de réaliser un Répertoire bibliographique universel, on se trouve plus loin que jamais, non seulement de l’unification des systèmes d’accès par sujets à laquelle tendait Paul Otlet, mais encore de la plus modeste coordination entre eux comme le proposait Gardin pour l’UNISIST”.
standardization, interoperability, the growing information system integration and the use of universal classifications in the web environments call for revisiting Paul Otlet’s valuable legacy. We devote our work for this task following on the footsteps of many researchers such as Boyd Rayward.

In our future we will look at the UDC use in the Web environments for a better evaluation of its capacity to handle multidimensionality. Examining this characteristic will definitely call for a comparison between the UDC and Ranganathan’s faceted classification and its use in the Web environments.

Another important investigation would be examining the role of the RIU and the contribution of P. Otlet’s knowledge organization theory to the Universal search.

References
Archives de l’Institut International de Bibliographie:


Kathryn La Barre - University of Illinois at Urbana-Champaign, USA
Rosa Inês de Novais Cordeiro - Universidade Federal Fluminense, Niterói, Brazil

Unmasking ‘That Obscure Object of Desire’:
A Brief Report from the Films and Facets project

Abstract
The use of facet analysis as an analytical tool for identifying salient access features of moving images was tested in this exploratory project. Twelve participants viewed three films, and created keywords and summaries for each film. These textual descriptions and access features from four moving image repositories were subject to facet analysis. A comparison of the facets elicited by this approach indicates that subject-related content, especially theme, genre and action may be highly salient access points for searchers in moving image repositories.

Overview
This exploratory project tested the use of facet and filmic analysis (CRG, 1957; La Barre, 2006, 2010; Ranganathan, 1933, 1937; Spiteri, 1998) as strategies for enhancing access and discovery of moving image resources. Developed jointly at the University of Illinois at Urbana-Champaign and Universidade Federal Fluminense in Rio de Janeiro, Brazil, the study assessed similarities and divergences in the content of written participant responses (n=12) to a set of three films across two different cultures (American and Brazilian). Participants were asked to identify salient film characteristics in the form of keywords and film summaries each person created after watching each film. The researchers conducted facet and filmic analysis of these written narratives as a way to uncover insights to support the creation of enhanced strategies for moving image representations. These insights could provide support for the creation of more robust access and discovery interfaces for films. One enhancement suggested by the findings provides support for the importance of indexing strategies for moving images that focus on subject-related attributes such as theme, action and genre.

Films and Facets Project
Today, the greatest challenge to the future of facet analysis is to determine the extent and manner in which this technique might be adopted and adapted for use in rapidly developing digital environments (La Barre, 2010, p. 268). The approach taken in this study integrates traditional facet analytic models as delineated by Vickery (1966, p. 43-44), and revisited by La Barre (2010) and replicates the facet analytical model developed in the project Folktales and Facets (La Barre & Tilley, 2010; La Barre & Tilley, 2012).

Films and Facets used the technique of facet analysis in conjunction with applicable principles of film study (Cordiero, 2000) to help identify important item features such as item format or theme, and the role of user characteristics such as language, experience, and understanding in textual descriptions created by study participants. The main objective of this study was to analyze written viewer responses to three films in order to extract salient aspects that might help improve retrieval and access to films in a variety of information environments (digital repositories, libraries and archives). This research sought to verify the utility of facet analysis and filmic analysis in the interpretation of the textual descriptions

1 The full results of this study will be presented in a forthcoming (2012) book chapter: That Obscure Object of Desire: Facets for Film Access and Discovery. In Diane Rasmussen Neal. (Ed.) Indexing and Retrieval of Non-Text Information. Berlin: De Gruyter Saur. The main focus of this report is upon the facet analytical aspect of this Films and Facets.
created by participants about selected films; and to observe any cross-cultural (national) similarities and differences in participant responses. Findings from this second objective will not be discussed in this brief report but initial observations are discussed in La Barre and Cordiero (2012).

Study Design
The following sections will discuss the approach taken by the researchers in selecting films for this study, analysis of keywords and summaries, and the protocol for facet analysis.

Film selection
This study included three films which varied in length (two short and one full length feature) at a total viewing time that would allow the participants to comfortably watch all three in one sitting in part to ensure completeness of data collection. Participants in this study constituted a sample of convenience, and were graduate students enrolled at each participating institution. Films for this study were selected along the following parameters:

a) **Availability**, in the public domain or via Creative Commons license
b) **Genre** – Fiction
c) **Film duration** (both short and full feature)
d) **Language of participants**: (English with Portuguese subtitles, or silent films)
e) **Image quality**: (no issues with viewability or sound quality)
f) **Cinematographic language**: one film was chosen from each category:

- **Classical narrative**: *A Corner in Wheat* (1909) Griffith – 13M 52s –silent. Social commentary (based on Frank Norris’ novel *The Pit*) in which the worlds of a poor farmer, and a wealthy, but ill-fated commodity market speculator collide.
- **Transition from classical to modern cinema**: *39 Steps* (1936) Hitchcock – sound 1H 32M. A well known, commercially produced feature-length British spy thriller featuring the wrongful accusation and eventual exoneration of the main character. Loosely based on the adventure novel, *The 39 Steps*, written by John Buchan.
- **Contemporary cinema**: *First Love* (2006) Flight and Camac – 6M 57s –silent. Amateur-produced film in the traditional style of a silent movie with intertitles and musical background material. Set in modern times, it is the story of a young boy who becomes infatuated with a young girl as he tries to earn money to buy her a gift (a sandwich for them to share).

Collection of summaries and keywords
After viewing each film participants were given an opportunity to take a break and then asked to write down a summary of the film and a set of words or phrases that « best described » each film. The following is an example of a summary and keyword written by the same participant for the short black and white film *First Love*:

*Summary*: First love is a short black and white silent film by Nick Flight and Scott Camac. The story involves a girl from modern day Tasmania, Australia who has a crush on the neighborhood girl. At first, his attempts to impress her are unsuccessful. He then devises a plan. Through perseverance and a few chores, he saves his money to purchase a sandwich for her. She is touched and shares the sandwich with him. (73 words)

*Keywords*: First love, first crush, childhood crush, young love, independent film, modern-day silent movie, short film

Analysis of participant summary and keywords.
After all responses were complete, Brazilian responses were translated into English by the Brazilian researcher and also entered into Google Translate (http://translate.google.com/). The textual descriptions were then analyzed and coded by each researcher. Both sets of coded textual descriptions were compared to identify whether each researcher had
identified reasonably similar terms or concepts or if there were areas of disagreement. There were no significant differences between coders and both sets were merged to create one set of coded responses.

A three-step protocol for the next phase of analysis was followed:

1. Terms and phrases from keywords were entered directly into a spreadsheet.
2. Concepts - in the form of representative terms and phrases - were drawn from the coded summaries and then entered into the spreadsheet. Keywords and summary concepts were not combined during this phase of analysis. Country of origin was coded for each participant.
3. Terms that were similar in meaning (e.g. crush and infatuation), or words existing in both plural and singular form (ball, balls) were combined in order to reduce duplication.

Table 1 illustrates the total wordcount of keywords and total concepts from movie summaries by country of participant. Table 2 compares average wordcount for the original textual descriptions as keywords and summaries across films. Films are listed in the order watched by participants.

Table 1: Comparative table of total assigned keywords (terms and phrases) and total concepts from movie summary descriptions by country (US –United States, BR – Brazil)

<table>
<thead>
<tr>
<th></th>
<th>Corner on Wheat</th>
<th>First Love</th>
<th>39 Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>US</td>
<td>BR</td>
<td>US</td>
</tr>
<tr>
<td>Total keywords (phrases / terms)</td>
<td>97</td>
<td>39</td>
<td>63</td>
</tr>
<tr>
<td>Total summary concepts</td>
<td>60</td>
<td>70</td>
<td>55</td>
</tr>
</tbody>
</table>

Table 2: compares the average wordcount for keywords and summaries across films

<table>
<thead>
<tr>
<th></th>
<th>Corner on Wheat</th>
<th>First Love</th>
<th>39 Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keywords : average wordcount</td>
<td>11</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>Summary : average wordcount</td>
<td>90</td>
<td>76</td>
<td>96</td>
</tr>
</tbody>
</table>

Table 3 shows the keywords that were most often selected by participants according to frequency of occurrence. For terms appearing in this table, the range of occurrence was between 8 (for those most frequently appearing) to 4 (the cut off point for inclusion in this Table. As expected, there was a fairly long tail of unique words, even after words were compared and combined for similarity.

Table 3: Top keywords by frequency of occurrence for the three films in this study

<table>
<thead>
<tr>
<th></th>
<th>Corner on Wheat</th>
<th>First Love</th>
<th>39 Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 Capitalism</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 Wheat</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black and White</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Love (First)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 Romance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Childlike</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Young Love</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hitchcock Spy</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Facet analysis

The process of facet analysis in this study began with the classic approach of asking a series of questions with regard to each textual description: What concept or concepts does it represent? In what conceptual category should this concept be included? What are the class relations between this concept and other concepts included in the same category? (La Barre, 2010; Vickery 1966). This classic approach « offers a set of principles and techniques that have now been applied in a variety of subject fields … It is potentially of considerable value to all prospective designers of retrieval systems in special fields.» (Vickery, 1960, p. 10).

This project follows models by Vickery (1966, p. 43-44) and revisited by La Barre and Tilley (2012). Steps included: clearly defining the subject field (in this case, moving image materials), and determining the user group of interest (graduate students interested in enhancing access to films). Moving from this standpoint, the next step sought to identify salient aspects of the three selected films that fit the user interests by consulting a representative set of materials - in this case three films, and a set of four moving image repositories which were analyzed to determine search and browsing features used to provide access to films. This provided a point of comparison between current practice and the project findings. The researchers assessed access points provided by four major archives of moving images: IMDb, BBC Motion Gallery, Moving Image Archive, and Petrobras Porta Curtas. Each provides free access to a minimum of 700 digital films through richly descriptive representations. The following table highlights the most common search or browsing features (or access points) for films on each site:

Table 1: Common access points offered for search and browse on four moving image archive sites

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Summary</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Date</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Title</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Associated persons</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Genre</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rights</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Length</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Color</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Format</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Keyword/subject</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Source</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
Findings

Facet analysis of the film keywords and summaries and the access features on the four portals resulted in a set of preliminary facets which are listed below. Not all of these facets are equivalent. Some are sub facets that link to more general or upper level facets. In the following list, all terms marked > denote higher level facets, under which the identified sub-facets may be located. Items marked with ** were not present in participant summaries or keywords but were identified through the analysis of the moving image repositories.

(Subject) Moving Images
>Attribute (of film)
   [Attributes sorted by production value]
   Color
   Length
   Format**
   Rights [access]
   Time [of production]
Association
   [types of associations]
   awards**
   derivations**
   collections**
   rating**
   persons [actors, director, screenwriter, etc.]
   locations [production, setting, studio]
   genre
Audience commentary
   [types of commentary]
   description
   emotion
   review
   rating
> Mise en scène²
   Character
   Object [item]
   Location [of action]
   Time [chronological of setting]
   Plot
   Action
   Setting /Location
   Time [chronology within film] [setting of film]
   Theme

Figures 1&2 (below) provide an introductory overview of the data by comparing the average distribution of facets across all three movies. Here theme, character, action, and to a somewhat lesser extent genre, rank as the most salient aspects of the three films. Figure 2 shows the distribution according to top level facets across movies. Here it is clear that participants found a variety of mise en scène features to be highly salient.

² everything that appears before the camera and its arrangement—composition, sets, props, actors, costumes, and lighting.
Fig. 1: Top facets assigned to keywords and summary concepts across movies.

Fig. 2: Average number of top facets assigned to keywords and summary concepts across movie

Conclusion
Facet analysis provided a useful set of analytical tools for understanding filmic responses in this project and gave a set of salient features to guide further phases of this project. A comparison of the facets elicited by this approach across films and across cultures indicates that subject-related content; especially theme, genre and action are highly salient access points for searchers in this group. This indicates a continued need to pay close attention to subject access in search and discovery systems. The findings of this project indicate that there is continued value in exploring ways to augment controlled vocabulary possibly by extending these with crowd-sourced subject terms from folksonomies as several moving image portals are now beginning attempt. With the increasing prevalence of digital multimedia resources, continued research is needed so that instead of the currently frustrating and dysfunctional relationship between ‘that obscure object of desire’ and those who seek to find moving images - a more user-aware paradigm might prevail.
References


Hemalata Iyer - University at Albany, State University of New York, USA
Amber D’Ambrosio - University at Albany, State University of New York, USA

Archetypes, Idealized Cognitive Models and Prototype Effect: A Discussion of Images and Cognition in Categorization

Abstract
Explores the usefulness of the idealized cognitive models (ICM) and related cognitive models in image organization. Discusses archetypes, ICMs and the prototypes visible in illustrations from historical children's literature and considers the pitfalls and challenges inherent in selecting prototypes of images for representation in Web-based resources, children’s catalogs and database.

Introduction
According to Jung, archetypes exist in a substratum reminiscent of Plato’s world of forms that is common to all humanity, the “collective unconscious”. Individuals build experiences and distinctive psychological characteristics based on the innate, imperceptible forms that they are born with. These archetypes emerge as symbols and patterns of behavior. These tend to manifest in images, art, myths, religion, etc. Examining these facets of human existence enables humans to indirectly identify the inherited archetypes. "The archetype is a tendency to form such representations of a motif - representations that can vary a great deal in detail without losing their basic pattern ... They are indeed an instinctive trend" (Jung, 1978, 58). One of the key ideas for knowledge organization is that they “can vary a great deal in detail without losing their basic pattern,” indicating that any number of variations on the central, core archetype can and do exist, but they all relate to, or cluster around, the archetype.

Archetypal images have played a significant role in folklore and literary studies, especially the works of Joseph Campbell and Northrop Frye. Work in these areas demonstrates how tracing archetypal images works across cultures. Joseph Campbell’s most famous work, The Hero with a Thousand Faces, offers a prime example of just how broadly the archetypal figure of the hero reaches. The Byronic hero, also sometimes known as a romantic hero as described by Frye, with all of the accompanying flaws offers an example of a variation on the archetype and is thought of as a literary archetype. The mother is perhaps the most powerful and natural archetypal figure to consider.

“The collective unconscious is every human’s biological inheritance of archetypes which offer potentials for meaning and image formation. An archetype can never be fully manifest in the conscious mind, let alone a literary text. What critics are really searching for are archetypal images, which will inevitably be affected by culture even if they bear the numinous stamp of the antecedent archetype. Jungian critics tend not to make a rigid distinction between the unrealizable archetype and the culturally influenced derivative, the archetypal image” (Rowland, 1972, 3). The archetypal tradition claims that humans function more basically in terms of symbolic images than literal concepts, which means they will be influenced more on a fundamental level by images, even those who function on a highly conceptual level. Particularly in the digital and Web environment, archetypes and their accompanying symbolic images offer the potential for organizing information in a way more naturally and universally appealing than that offered by conceptual language.
This paper is less concerned with the numerous potential causes of archetypal images or the possibility of transcendental archetypes and their resemblance to Platonic forms that muddy the waters of Jungian thought than it is with the basic idea that these types of symbolic images constitute a knowledge structure that has the potential to shape how we organize knowledge in the Web environment. Whether or not someone believes in the “collective unconscious” and its power, archetypal images can be utilized as prototype images in knowledge organization to group concepts under a broader idea represented by the image. Images speak across cultures in a manner that language cannot and can also span generations and age groups. Visual memory also often operates much more effectively than conceptual or word memory. While the applications for digital image collections are readily apparent, this can also be used when designing information systems aimed at children or for designing cross-cultural or cross-linguistic information systems. Representative images are also used in browsing large image collections.

Prototypes resemble archetypal images. Viewing this from a cognitive perspective, the prototype effect occurs within an idealized cognitive model. Archetypes exist outside human comprehension, but the archetypal images reveal aspects of the archetype but with the effects of the cultural context. An idealized cognitive model (ICM) depends on context, on the conceptual maps of a society, and the prototypes thus generated cluster and relate to each other in our minds. Hence, the ICMs are seen as deeper structures and the prototypes as more of a surface-level phenomenon. Archetypes are somewhere outside of humans but shared by all humans on some unconscious level and might be thought of as the root source of idealized cognitive models while the social and cultural context determines all other aspects of the model. However, it is that vital crux where archetype becomes an ICM through the window of context that makes using images potentially as challenging as promising for the organization of information. This paper further discusses archetypes, ICMs and the prototypes visible in illustrations from historical children’s literature and considers the pitfalls and challenges inherent in selecting prototypes for representation in Web-based resources, children’s catalogs and databases.

Objectives

The objective is to explore the usefulness of ICM and related cognitive models to image organization and to examine the illustrations in children’s literature from this perspective.

Methodology

The historical children’s literature collection held by the M. E. Grenander Department of Special Collections and Archives at the University at Albany, State University of New York was used for this study. Fifty books that portray blacks in their illustrations were selected randomly from the collection. The illustrations were analyzed in terms of gender representation, activities engaged in, occupations, clothing, inter-relations portrayed and background scenes.

Cognitive Models

Cognition and categorization have a pivotal role in knowledge organization. How do we comprehend categories and what are the most relevant cognitive models that come to play in the process of understanding and category formation? The concept of metonymies, paragons, typical examples, and stereotypes are seen as ICMs that produce a prototype effect. A large part of our knowledge is organized based on typical examples, and we reason, draw inferences and create generalizations based on these examples. We tend to use the typical cases to generalize about the non-typical ones and not vice versa. This leads to what George Lakoff calls the prototype effect. Stereotypes are similar to typical examples but slightly different, as they tend to be oversimplified and often carry negative
connotations while paragons and ideals have made a positive impression to the degree that people will seek to emulate them. *Metonymies* are yet another source of prototype effect. A metonym is a part that takes the place of the whole; the part comes to substitute for or represent the whole (Iyer, 1995; Lakoff, 1987).

Lakoff’s classic example of “mother” presents a *cluster model* that entails a clustering of several representations/cognitive models of the mother idea is very relevant in this context; an additional level of prototype effect occurs in this category. The social stereotype of a housewife mother frequently stands for the mother category as a whole and defines the cultural expectations of what a mother is supposed to be. This is associated with caring for and raising children and additionally entails the nurturance model of the mother concept. Thereby it serves as a basis of judgment for other models of the mother category and produces another layer of prototype effect. The resulting clustering of cognitive models of the mother include the housewife model in the center with the various other models, such as adoptive mother, surrogate mother, working mother, foster mother, donor mothers etc., representing the birth model, genealogical model, marital model and so on, converging and forming a composite mother category and a *composite prototype* with attributes from the various models. The other models are comprehended in contrast with the nurturance model, and this he calls as the *radial structure* and the presence of such a structure within a category can be a the source of prototype effects (Iyer, 1995, 51-52; Lakoff, 1987, 80-84).

Theoretically speaking, this has no place in the classical theory of categorization in which all members of a category possess the necessary and sufficient conditions to belong to a category. Stereotypes have no place in such an approach to conceptual structuring. However, in reality, stereotypes do have a significant role in characterizing concepts, and they define the normal expectation, and this plays an important role in cognition, like in the mother example where the other models of mother, such as working mothers, are defined in contrast to the nurturance models. Thus we need to take cognizance of *metonymies, stereotypes, typical examples*, and *paragons and ideals* that shape and influence our thinking. This is especially true when we are organizing collections, both texts and images, wherein it is often necessary to choose representative examples for categories, such as in the choice of thumbnails used for browsing image collections, when representing categories of information for children or endeavoring to create cross-linguistic collections. In this paper the characterization of blacks in historical children’s literature is examined and discussed from the cognitive point of view as a means of understanding the possibilities and important variables to consider when selecting prototype images.

**Observation and Discussion of the Illustrations**

Gender inequity was evident; boys were represented more frequently than girls; Men were shown as working very hard in the plantations and on their own farms. In such illustrations women were in the background and girls were infrequently represented and, even if they were, they were seen in the background. LaDow in the content analysis study of randomly selected picture books from the Mishawaka, Indiana Public Library also observed that in the illustrations males appeared four times more than the females (LaDow, 1976). Women were shown with a matronly bearing, engaged in preparing a meal, often coaxing and disciplining a child at the dining table. Some illustrations represented them outdoors with a basket full of fruits that they had picked on the farm. Gender stereotyping is evident in the illustrations. Children were seen playing, chatting, sitting on the farm fence, riding in a tyre, dancing and helping the adults in their outdoor activities and work. They appear as being very joyful in the outdoor settings. Most often they are seen with their pets, especially puppies. The books set in the south especially portray plantations with adults and children drawing a wagon full of produce. Stereotyping with reference to gender,
racial features and occupational roles, was observed. The surroundings are usually modest log cabin homes.

As regards archetypes, those of caregiver, mother, outsider, hero, friend and the master-servant dynamic seem to be present in the illustrations. The archetypes observed in the illustrations overlap with the findings of Beth Jones Ricks in her study of Newbery children’s literature. Ricks identified 12 archetypes among the female characters across all genres, thereby demonstrating that archetypes are universal and are experienced by all regardless of race, religion, time period and gender. Young female characters are friends, caregivers, orphans, innocents, and seekers; warriors and fools are found in both young and old characters; older women who are in charge are presented as evil, tyrannical and untrustworthy (Ricks, 2004). Specific instances are discussed in the following section.

**Challenges in Choosing Image Prototypes**

When considering the possibility of using prototype images for databases or websites for children, for example, many challenges exist in choosing images. The encroaching threat to ICMs and prototypes are some literary archetypes, such as the wicked stepmother, and stereotypes, which bear resemblances to these means of classifying knowledge, but include characteristics, usually negative, not necessary for organization into conceptual clusters. Children’s literature might seem an ideal location from which to draw prototype images for child-oriented databases and Web resources, but examining illustrations from historical children’s literature demonstrates the danger for stereotypical representations to leak into illustrations. Looking at children’s literature related to an ethnic group frequently stereotyped over an extended period of time in the United States reveals the challenges in choosing prototype images that do not bear unnecessary “baggage” while standing in for an ICM. Perhaps the most eerily representative image of what a prototype should avoid comes from a children’s book simply titled *George Washington* by Ingri and Edgar Parin D’Aulaire (copyright 1936, published 1940). The illustration depicts the wedding ball of George and Martha Washington through a window with black children and a dog in the foreground peering in on a lavish scene in which they have no place. A prototype should leave no one feeling like an outsider looking in through glass, though this may be seen as reminiscent of the outsider archetype.

The majority of the books examined used hand-drawn illustrations to depict characters and events in the books, though two books relied on photographs. One of the books using photographs instead of illustrations is *Tobe* by Stella Gentry Sharpe with photographs by Charles Farrell, published in 1939. Like many of the illustrated books, this book depicts characters engaged in farming activities, but the differences in details between illustrations and photographs are notable. The photographs show the family as it actually is without the details that make most of the illustrated characters into caricatures that reflect negative stereotypes as is the case with a book like *Little Brown Koko’s Pets and Playmates* by Blanche Seale Hunt (1959), which also takes place on a farm, and has illustrations of all its black characters with large red lips. The mother from *Tobe* is an average-sized woman sewing and listening to the radio as she sits with her husband or shown walking into church at the head of her family; the mother from *Koko* is drawn large, wearing an apron and head cloth, and in the kitchen cooking or in a rocking chair knitting. Both of these books show a mother figure, but they depict the mother figure in distinctly different ways. The photograph of Tobe’s mother sewing compared with the most compositionally similar image of Koko’s mother knitting emphasizes the black mother caricature of Koko’s mother. Other mother or nanny figures drawn in similar fashions appear in other Afro-American children’s literature. Tobe’s mother from the photographs might be considered as representative of a cluster model ICM, portrayed as possessing more facets and without the negative elements of a stereotyped caricature. This demonstrates the necessity of choosing prototype images with care because something as seemingly innocent as a woman
sewing/knitting can have so much more meaning because of the stereotypes that exist in a society’s conceptual map at any given time and bleed so easily into illustrations.

Attempting to classify activities using a prototype image drawn from children’s literature might also prove challenging. The children’s literature examined typically takes place on a farm or shows a plantation with black men harvesting crops. Even a category like “occupation” can be complicated by an image of black people as laborers. The other book illustrated through the use of photographs once again proves the best counter-example to the majority of the children’s literature available. *My Dog Rinty* by Ellen Tarry and Marie Hall Ets (1946) is set in Harlem, New York and shows black characters in an urban setting. Compared to books set in the countryside, there is a well-dressed newspaper editor in *Rinty* and at one point the father to the family, also wearing a suit, reads to his children from a library book. The mother is shown in the kitchen cooking and accepting a note from her son, a neighbor is shown bathing her baby, and other women in the book hold professional positions as store clerks, nurses or receptionists. When choosing a prototype image for “occupation”, should it be a man or a woman depicted? What ethnicity should they be? What type of occupation would avoid issues of bias? In the books described above, there are people in a variety of occupations but which gender in which occupation would be most neutral?

The children’s literature also allows for consideration of paragons. A folk hero is a type of paragon, and John Henry is the prime example of a black folk hero. *John Henry and the Double-Jointed Steam Drill* by Irwin Shapiro (1945) is a children’s book in the collection about this famous folk hero. However, the visual depictions of Henry and other black characters in the book convey a different impression. The muscles are over-emphasized, the limbs and fingers too long, the eyes often blank, the lips too large, and the characters frequently hunched over in an awkward manner. Illustrating John Henry as larger-than-life would hold with the folk hero tradition, but there is a wide-variety of other details in the illustrations that make this depiction of a famous folk hero troubling. While it might be ascribed to the illustrator’s individual style, stereotypes of black men could play a role in the skewing of the images in this children’s book. Who chooses to create an image and how they feel about the subject of that image can impact what a viewer experiences regardless of how positive the subject of an image is might be. Because of the inherently superlative aspects of a paragon or ideal it would seem easy to choose the resultant prototype image to represent a category, but, as this example shows, a great deal can be read from or into an image that may undermine its status as prototype.

**Conclusion**

Because prototype images stand in for a group of related concepts, extreme care must be taken in choosing them. A word or phrase can have multiple connotations but ultimately lacks the details of an image that can have far more meanings and more far-reaching consequences. When considering databases or catalogs geared toward children, the psychological effects that images can have need to be taken into consideration. A great deal of work has been done on the impact of stereotypes in US children’s literature on a child’s psyche, particularly for children from under-represented groups (Roethler, 1998). In terms of cross-linguistic and cross-cultural efforts in categorization, such as might be undertaken on the Web, the vastly different contexts make choosing neutral images challenging. The conceptual maps of societies differ from each other, so while images might circumvent issues of language, they can result in vastly different interpretations depending on the culture encountering the image. Even over time conceptual maps change, which can impact archetypal images even if at the root there lays an unchanging ideal. A prime example of this would be the literary archetype of the hero. The hero of Greek mythology is not the
knight of Arthurian legend or the misanthropic hero of Shakespeare and Byron or the American folk hero or the super hero of comic books in the United States in spite of all of these heroes originating in western European cultural traditions. No matter how information professionals engage with the organization of knowledge, whether in libraries or on the Web, the fluidity of human experience requires constant change in approach and methods along with a persistent awareness that one person’s interpretation will not be another’s.

Moving forward in information organization for the Web, children’s databases and other categorization methods that might benefit from prototype images, there are several important factors to keep in mind from the discussion above. As explored extensively in this paper, there are dangers of caricature or negative stereotyping appearing in images, but this should not prevent the use of prototype images. There are ways to approach this and minimize excluding or marginalizing the people encountering prototype images. One way would be to use a variety of prototype images, particularly those that result from the clustering effect described by Lakoff. An image that captures the multi-faceted nature of a category while still representing that category would acknowledge the wide-range of cognitive models that converge around a concept. A selection of images, either rotating through or as a small collage, could also be used. Instead of the category of “occupation” being represented by a single image of a CEO or a doctor, images of a farmer, a mechanic, a musician, an artist, etc. could be provided. It is likely that broad categories like this will be most difficult to represent with an unbiased prototype image. Narrower categories run more of a risk of stereotyping rather than marginalizing or excluding a person or group of people in a prototype image. Having a large and diverse group of people involved in image selection would also help to ensure that the most neutral prototype images are chosen. Regardless of the method chosen, the prototype image that results from an ICM must always be acknowledged to arise from a particular social and cultural context and steps taken to recognize elements of an image that might be considered negative, exclusionary or marginalizing.

References
Aboutness in Fiction:
Methodological Perspectives for Knowledge Organization

Abstract
The subject analysis of narrative texts of fiction is complex; the methodological model of identification of concepts as elaborated for scientific texts is not applicable to fiction. It is proposed here that theoretical and methodological use of the Generative Trajectory of Meaning postulated by Greimas may contribute to the identification of aboutness in narrative texts of fiction.

Context
Subject analysis in knowledge organization reveals different levels of complexity depending on the nature of the material. The complexity is significantly higher when dealing with literature and fiction. The traditional methodological model of subject analysis (specially the concepts identification procedures) elaborated for scientific texts does not seem to fit to narrative texts, since these materials have different characteristics compared with scientific documents in terms of both structure and content. There is also a lack of methodological studies on subject analysis of fiction compared with the scientific literature; Information Science field, apparently, does not give necessary attention to fiction considering them primarily as material for leisure and entertainment. There are texts directed only towards entertainment, the so called pulp fiction; however, there are those which are considered works of art with words, the foundation of literature.

There are institutions in which fiction occupies primacy, for example, the public libraries, as well as university libraries with focus on arts, literature and social sciences. It must also be said that the development of reading habits, particularly in Brazil, begins with the reading of literature or literary texts; because of this there is a need for greater attention to issues related to organization of these.

Classification and indexing of fiction is not an issue that has only been discussed in the last few decades. As a matter of fact, presumably, the first article on the subject appeared in 1898, when the noteworthy British librarian Ernest Baker wrote about the classification of fiction in the Library World, and in the beginning of the 20th century a discussion concerning the same issue took place in the American Library Association (ALA). (Eriksson, 2005, p. 01). In the last thirty years, the subject of fiction seems to have come up again, and several studies have been reported; Peitersen (1978, 1979, 1983, 1984,1998), Beghtol(1986, 1992, 1994, 1995, 1997), Hayes(1992), Nielsen(1997), Saartit(1999), García-Marco; García-Marco(1997), Moraes, J. B. E.; Guimarães, J.A.C.(2006), Guimarães, J. A. C.; Moraes, J. B. E.; Guarido, M. D. M.( 2007), Moraes, J. B. E.(2008), García-Marco et. al. (2010) among others.

The problem that motivated this paper is the lack of an analytical procedure to guide the identification of aboutness of fiction. As a hypothesis, we present the methodological tool offered by the Greimas’ semiotics theory called Generative Trajectory of Meaning. As a general objective, this study intends to make methodological contributions to the area of knowledge organization in imaginative literature or fiction. More specifically, it aims a) to study Generative Trajectory of Meaning as a tool to organize literature or fiction, b) to discuss the concept of aboutness; and c) to analyze discursive semantics as a tool to study imaginative literature or fiction.
Aboutness

John Hutchins (1977) was the first one to discuss the *aboutness* question, according to Caffo (1988, 24). For Hutchins, the crucial problem is to identify "what documents are about". In author's conception, the first question to be made is "what is meant by the topic of a document in the context of an information system?" For Hutchins (1977, 17) a superficial answer would be: the topic of a document is the subject description in an index entry relating to that document, but, in fact, there is rarely a straight equation of subject description and 'what the document is about'. As can be seen in post-coordinate indexing systems the subject description in an index entry may represent only part of the document's content.

A second question by Hutchins (1977, 17) is: "What do we mean by the content of a document?" According to the author, we need to be clear about the distinction between the 'sense' of a linguistic expression and the 'reference' of that expression. In other words, the choice of a particular expression to refer to a particular individual is determined by the appropriateness of the expression, and whether an expression is appropriate or not depends primarily on its meaning or 'sense'. The sense of an expression is determined principally by its relationship to other expressions of the language (Hutchins, 1977, 18).

Maron (1977) discusses the main point of a retrieval system and observes: "Since *about* is at the heart of indexing, how are we to formulate any proper theory of indexing if we cannot explicate precisely the key concept of *about*?" (Maron, 1977, p. 38).

Thus, Maron (1977, 40) suggests that there are several different *about* concepts. The first one is the *S-about* (subjective about), a relationship between a document and the resulting inner experience of its readers, and cannot be analyzed further in objective terms.; the *O-about* (objective about), the interpretation for this concept of *about* is obtained by considering an external or observer’s point of view, as opposed to the internal or subjective point of view, and refers to the (actual or potential) behavior of asking or searching for writings; *R-about* (retrieval about), explained statistically and behaviorally, in terms of those who would be satisfied with it and where such people would be looking for it.

Despite the complexity of the definitions, it's interesting to note that the author considers not only the document itself, but also the search and retrieval contexts.

Ingwersen (1992) also studied the *aboutness* question. According to him, *aboutness* can be seen from four different points of view: from the *author* – this the common method of natural language representation; from the *indexer* – this implies in transforming author's natural language in a documental language accepted or created by a indexer; from the *user* – formulated from the new knowledge user's needs; from the *request aboutness* – formulated by the user.

Caffo (1988) presents another view of *aboutness*, considering the relationship between the "document discourse towards a concept and the concept theoretically defined as a part of the general knowledge" (Caffo, 1988, 23). Indexing is not concerned so much with the conceptual analysis in knowledge theoretical complex, but with the conceptual dimension of a materialized knowledge, in this case the document, as Michel Buckland (1991), among others, recognize. Beghtol (1986) distinguishes 'meaning' and 'aboutness':

Whatever terms are chosen, a distinction between 'aboutness' and 'meaning', as the terms used here, seems justifiable on the assumption that a document has an intrinsic subject, an 'aboutness', that is at least to some extent independent of the temporary usage to which an individual might put one or more of its meanings. (Beghtol, 1986, p. 85).
As can be seen, Beghtol (1986, 85) refers to *aboutness* itself and to a *meaning*. Thus, according to the author, *aboutness* is something intrinsic to the document, with a relatively permanent nature, itself unchanged; on the other hand, we have a *meaning* that can change according to the situation, location, user's interests, etc. In other words, "the same document can have different meanings for the same reader at different times, but the document itself, is assumed to possess a fundamental aboutness" (Beghtol, 1986, 85).

Finally, we have to mention Layne (2002) who uses Panofsky’s classification scheme to deal with "subject access to art images". The author uses the expressions *aboutness* and *aboutless* in order to search for the contents of art images. As we deal only with imaginative fiction, we are not going to explore Layne’s theory.

**Generative Trajectory of Meaning**

For Greimas; Courtés (2008) the *Generative Trajectory of Meaning* can be defined as follows:

> We designate by the expression generative trajectory the general economy of a semiotic theory (or just linguistic), as say, the disposition of its components one related to the others, and this in a generative perspective, e.g., postulating that every semiotic object can be defined according the way of its production, the components that intermediate in this process are articulated to each other according a ‘trajectory’ from the simplest to the most complex, and from the most abstract to the most concrete. (Greimas; Courtés, 2008, p. 232).

<table>
<thead>
<tr>
<th>Generative Trajectory of Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semio – narrative structures</td>
</tr>
<tr>
<td>Syntactic component</td>
</tr>
<tr>
<td>deep level</td>
</tr>
<tr>
<td>Fundamental Syntax</td>
</tr>
<tr>
<td>Surface level</td>
</tr>
<tr>
<td>Surface Narrative Syntax</td>
</tr>
<tr>
<td>Semantic component</td>
</tr>
<tr>
<td>Fundamental Semantics</td>
</tr>
<tr>
<td>Narrative Semantics</td>
</tr>
<tr>
<td>Discoursive structures</td>
</tr>
<tr>
<td>Discoursive Syntax</td>
</tr>
<tr>
<td>Discoursivization (actorialization, temporalization, spationalization)</td>
</tr>
<tr>
<td>Discoursive Semantics</td>
</tr>
<tr>
<td>Thematization Figurativization</td>
</tr>
</tbody>
</table>

Source: Greimas; Courtés, 2008, p. 235

Fiorin (2011, 17) says that the trajectory scheme is composed by:

- Semio-narrative structures with syntactic components (deep level: fundamental syntax; surface level: narrative syntax) and semantic components (fundamental semantics; and narrative semantics) and
- Discourse structures with syntactic components (discoursive syntax; discoursivization – actorialization, temporalization, spationalization) and semantic component (discoursive semantics; thematization; figurativization)

Moraes & Guimarães (2006, 39) claim that the generative trajectory of meaning is composed of three levels: *fundamental structures; narrative structures and discursive structures*. Each one of them is detailed by Moraes; Guimarães; Guarido (2007):

In *fundamental structures* we can find the semantic categories that ordinate the text content in a more general and abstract way; the *narrative structures* can be defined as state changes in terms of conjunction and disjunction (manipulation, competence, performance and sanction); in the *discursive structure level* emerge the abstract narrative structures that can be achieved by figurativization or thematization, e.g., by themes or figures (Moraes; Guimarães; Guarido, 2007, p. 97).
Deep or Fundamental Level

Moraes & Guimarães (2006, p. 39) say in fundamental structures are found the semantic categories that ordinate the text contents in an abstract. To understand the value of each category it is necessary to find elements in the other concrete text levels.

According to Fiorin (2011, 20) we find in the deep level the semantic categories that are the basis of text building, and these categories are based in a difference, in an opposition. Greimas called this opposition Semiotic Square:

Semiotic Square can be understood as a visual representation of a logical articulation from any semantic category. The elementary meaning structure, when defined – in a first moment – as a relationship at least with two terms, lays only on a distinction of opposition which characterizes the language paradigmatical axis: it is, therefore, enough to built a paradigm composed by \( n \) terms, but do not allow, by this, distinguish, in this paradigm, semantic categories based on isotopie ("kinship") of the distinguished that can be recognized. (Greimas & Courtés, 2008, p. 400).

In this way, the deep level syntax involves two operations: assertion and negation. These transforming operations, which occur in the narrative and discursive levels can be represented by the semantic square, as shown in Figure 1:

![Fig. 1: Semiotic Square](source: Greimas; Courtés, 2008, p. 402.)

Negation and assertion operations are in the surface and other articulations are in the deep level. Negation is close to contradictory terms that many times work as passage terms. Assertion is linked to contradictory terms that articulate the opposition that exists in a text.

Narrative Level

For Moraes & Guimarães (2006, 39) the narrative structures can be defined as state changes in terms of conjunction and disjunction.

Texts are not minimal narratives, but complex ones where the state utters (making and being) are hierarchically organized in a canonical structure composed by four phases: manipulation component has to do with changes in wanting-to-do and/or having-to-do; competence is the component that has to do with changes (creation, maintenance, increase, decrease, loss) in the prerequisite elements of performance (accomplishing the action); performance is the component that has to do with the action's realization (in the strict sense), which was made possible by positive competence. Performance presupposes competence (and manipulation, naturally, since it involves wanting-to-do and having-to-do, just as competence does): If there is performance, then competence was necessarily positive; sanction is the component that concerns the epistemic judgment (evaluation) of performance and the accompanying retribution that the performing subject has incurred.

Moraes; Guimarães; Guarido (2007, p. 97) say that a canonical order does not mean a preformatted order in which all narrative texts must to fit, but, on the contrary, some phases can only be assumed, or have a different detach than others.
Discoursive Level

At the Discoursive Level the narrative level abstract forms are covered by terms which give them a concrete form. In this sense, we have two levels that compound the concretization of the narrative structures: thematization and figurativization.

According to Fiorin (2011, 72) figures apply to natural world elements, real or fictional, that can be perceived by senses: stone, boy, moon light, fly. Themes apply to elements that organize, categorize and regulate perceived reality: hostility, ingenuity, to imagine, happy.

The opposition theme/figure can lead to the conclusion that it is an opposition abstract/concrete. However, one must have in mind that concrete and abstract are not polar terms which are opposite each other, but they constitute a continuum that goes, in a gradual way, from most abstract to most concrete.

Themes are organized through a chaining called thematic trajectory. The same thing happens with figures, and this chaining is called figurative trajectory.

In fact, the themes array, concatenated in the body of the narrative texts, will guarantee the themes extraction. In the same way the figures are chained in a coherent way in order to have a text harmony. If there is a break in the internal coherence of themes and figures web the text may become improbable or new levels of meaning may arise. Thus, according to Fiorin (2011, 79) everything in a text is relationship. For this, to find the meaning of a figure set chained is to find the theme underlying it.

A meaning amalgamator element for the thematic and figurative trajectory is the isotopy. According to Greimas; Courtés (2008, 275-8):

A.J. Greimas has taken to the physics-chemistry domain the term isotopy and transferred it to the semantic analysis, gave it a specific meaning, having in consideration its new application field. (...) Isotopy constitutes a new way of reading that become homogeneous the text surface, since it allow eliminate ambiguities. One can say, therefore, that arise isotopies is identify the semantic continuities which makes the read text in a coherent set. (Greimas; Courtés, 2008, 275-8).

Generative Trajectory of Meaning and Aboutness

The scientific texts seem to possess an apparently „stable” structure, independent of the subject field, that seems to facilitate the work of the indexer. Therefore, it has something similar to a script to be followed, which, most of the times, starts by the title, the keywords, and the abstract.

On the other hand, fiction also can possess some structural elements that can be considered constant – although very different from the scientific texts - which can facilitate the performance of the indexer when searching for aboutness of these texts.

In this way, there is a superstructure, which presumes that in fiction, mainly in narrative texts, action is predominant, as well a sequence of events in which complication, evaluation and resolution are involved.

Moreover, the indexer also can use the narrative structures to locate the manipulation, competence, performance, cognitive sanction and pragmatic sanction, constituent elements of the call canonical sequence of the narrative structures, according to the Greimas’ Generative Sense Course.

An approximation of superstructure theory, postulated by Van Dijk (1997), and the narrative structures, postulated by Greimas shows that both can be useful in a complementary way as tools to the subject indexer when searching the aboutness of fiction literature. The first one provides concrete and consistent elements in order to identify the form of the text, confirming it as a narrative text; the second one allows to the analyst identify some parts which are more or less constants in this kind of text. However not all the canonic sequence elements appears in a text, some parts can be presumed, in general the initial portion, manipulation, and the final, sanction, are present during the analysis.
It must be mentioned the *macrostructures*, also postulated by Van Dijk (1997), as an element of fiction analysis. In that, the presence of characters, human or anthropomorphized, as well an idea of action, of state changing, of transformation or happenings, in which the chronological sequence is fundamental, can be identified.

The analysis of *fundamental structures*, postulated by Greimas, can also be linked to *aboutness*. The first concept is concerned with the semantic question which originates the process of signification, e.g., the primary and fundamental element, and that builds its relationships in terms of values. The second concept is concerned with that basic element over which the text is built, and can be observed in terms of relationship, so at this point the *meaning* can be established. The focus, in this case, is the process of insertion of a document meaning in a system which serves information retrieval.

**Conclusion**

However there are some differences in the origin fields; we pointed out that the studies on *aboutness* and *generative meaning trajectory* have some points in common. Of course, none of them is linked to the coincidence of the term *meaning*. It’s the same term, but concepts are much different.

But we can consider the concept of *fundamental structures* so much close to the *aboutness*; the first one considering an opposition between two terms indicated by the semantic square; the second one indicates the text intrinsic content.

In fact, many studies carried on with short stories (Moraes; Guimarães, 2006, Guimarães; Moraes; Guarido, 2007) confirmed that the Generative Sense Course can be useful to find the *aboutness* of imaginative literature or fiction, and opens the possibility to future studies on larger texts, like romances.

**References**


Alignment of Conceptual Structures in Controlled Vocabularies in the Domain of Chinese Art – A Discussion of Issues and Patterns

Abstract
Based on our recent sub-project of the Chinese AAT-Taiwan Project, this paper reports issues regarding the alignment of the controlled vocabularies in the domain of Chinese art. The conceptual structures of the concepts for Chinese art in the National Palace Museum (NPM) Vocabularies and the Art & Architecture Thesaurus (AAT) are studied and patterns were identified in the effort of achieving semantic interoperability. The findings presented in the paper are meaningful to the research on the semantic interoperability of multilingual KOS, especially when dealing with cultural-related concepts that cannot be exactly aligned in vocabularies due to the discrepancies in the conceptual structures.

Introduction
Semantic interoperability of knowledge organization systems (KOS) is one of the important issues discussed in literature related to information services at all levels due to its direct impact on the quality of searching and browsing performed by users. Semantic interoperability is the capacity for different agents, services, and applications to communicate data, information, and knowledge while ensuring accuracy and preserving the meaning of that data, information, and knowledge (Zeng and Chan, 2010). In any effort that involves multilingual KOS, the language and cultural variants bring a great degree of complexity, especially when perspectives of different cultures need to be integrated. This paper attempts to explore, within the domain of Chinese art, the interoperability of multilingual/multi-cultural vocabularies by analyzing the issues of aligning the conceptual structures of Chinese and English terminology in the controlled vocabularies observed in our recent sub-project of the Chinese AAT-Taiwan Project (2010- ). Here –conceptual structure” means the semantic context of a concept such as the hierarchical structure or category a concept is situated in a thesaurus. Although in the project we also completed the major mapping process for the Chinese and English terms representing concepts (Chen, Wu, Peng, & Chang, 2010), this paper will only address the issues related to the conceptual structures of these concepts for Chinese art in the National Palace Museum (NPM) Vocabularies and Art & Architecture Thesaurus (AAT). The ultimate goal of the paper is to explore the extent of semantic interoperability of Chinese art-related terminology in current Western art-based thesauri, as reflected particularly through the discrepancies (i.e., the state of disagreeing or being at variance) in the conceptual structures of concepts in the NPM Vocabularies and AAT.

Background of the vocabularies
As one of the top five museums in the world, the National Palace Museum (NPM) in Taipei holds a large collection of the most precious items of Chinese art. In the registration and description of its over 80,000 items, NPM implements its Metadata Requirement Specifications. The source vocabulary used in this research is the whole set of controlled vocabularies (over 2000 Chinese terms representing concepts) from the NPM’s Metadata
Requirement Specifications (Academia Sinica and National Palace Museum, 2001, 2004) (referred to as NPM Vocabularies). Controlled values are provided for each of the metadata elements in the Specifications. For example, for the NPM element “Function” of artifacts, there are pre-defined values such as *drinking vessels, costumes, and musical instruments*; for the element “Technique” regarding calligraphy and painting the values include *baimiao* (a technique executed in fine lines and delicate colors), *gonbi* (a technique using meticulous brush work with attention to details, forms and standards), *xieyi* (a method using free sketch or freehand brushwork), etc. NPM did not adopt a thesaurus standard to organize these values. The values defined for different attributes are grouped into a set of controlled vocabularies that are structured with one to four hierarchies. The grouping and hierarchical structure setting are based on the long-time research experiences and best practices of the Chinese art researchers. These controlled vocabularies have been used to index and describe 80,000 items in the Museum’s collection.

*AAT* is a multilingual thesaurus developed and maintained by the Getty Research Institute in the United States. Currently it has over 35,000 concepts, each represented by English and other language terms (Spanish and Dutch terms for each concept, French and Italian terms for most of the concepts, and increasingly added Chinese and German terms). *AAT* began with Western culture-based content and has expanded continually to include other cultures around the world. *AAT* is constructed according to the ISO standard for thesauri (*ISO 5964, 1985*), always considering semantic relationships (hierarchical and associative relationships) between concepts and equivalence relationships of terms representing the same concept (Harpring, 2010). The hierarchical relationships include genus/species, whole/part, and class/instance relationships between and/or among concepts. *AAT* organizes all concepts within seven facets: Associated Concepts, Physical Attributes, Styles and Periods, Agents, Activities, Materials, and Objects (*Art & Architecture Thesaurus Online, 2000*). There are well-established hierarchies for all concepts in the thesaurus. For example, in the Style and Periods Facet concepts are grouped according to general era (ancient, prehistoric, pre-ceramic, etc.), region (Asian, Early Western World, etc.), and generic styles and periods (antique, contemporary, southwest, etc.). All Chinese painting styles are organized under the guide term <Chinese painting styles> and a total of 17 concepts (including *baimiao*) are included. *Gongbi*, on the other hand, is grouped under the guide term <painting techniques by application method or circumstances>, which is in the Activities Facet.

**Research questions**

When trying to align the NPM Vocabularies with *AAT*, like in any other multilingual vocabulary mapping process, we first dealt with various types of equivalence relationships between terms representing the concepts. These include exact equivalence and partial equivalence. In addition, we also faced with issues in aligning the terminology’s conceptual structures presented in the vocabularies. Taking an example of a concept *baimiao*, it is listed under “Technique” in the NPM Vocabularies; its English equivalent is, however, categorized as a “style” under the “Styles and Periods Facet” in *AAT*. This would raise a question in the alignment: should one of the conceptual structures be prioritized over the other? If we chose *AAT* as the principle structure, the conceptual structure of “Technique” in NPM would be disorganized, which may result in losing its integrity in presenting the knowledge of Chinese art. Meanwhile, if we would not keep up with the conceptual structure defined by *AAT*, then the whole point of aligning with *AAT* would be minimized. Then, should we keep both conceptual structures? If so, how are we going to process and achieve the alignment? In fact, this is a widely found problem. Thinking of many specialized terms in the domain of Chinese culture and art, we would ask the similar
questions again and again. Eventually, could each of the Chinese art concepts be located in an equivalent or similar conceptual structure in AAT? Can all of these concepts be mapped reasonably to AAT? Also, for those Chinese and English terms that met the “exact equivalence” mapping criteria, are their conceptual structures different? Furthermore, are there different levels of discrepancies in their contexts?

This study attempts to explore such situations while addressing the general research questions related to the structural discrepancies in the alignment of terminology in multilingual and multi-cultural context, the essence of the issues underneath the situations, and the possible solutions.

Methodology

The study considered the metadata elements to which the controlled values belong as the basic units. This resulted in units such as “Technique (P02)”, “Material (A06)”, and “Archaeological Culture (A02)”. Consequently the study examined the controlled values listed under each metadata element and identified those that have strong Chinese art and culture characteristics. For instance, unit “Archaeological Culture (A02)” refers to Chinese archaeological culture, under which the 13 values, (e.g., Yangshao Culture) are all unique to the Chinese culture. On the contrary, unit “Material (A06)” refers to the material of artifacts, under which there are 110 controlled values, mostly not being Chinese culture-dependent (e.g., shell (A06002), bamboo (A06030), and emerald (A06099)). It is also true that a unit may encompass both unique Chinese art terminology and culturally shared terminology. For instance, unit “Technique (P02)” refers to calligraphy and painting techniques, under which the 35 controlled values are mostly unique to Chinese art (e.g., baimiao (P02001), gongbi (P02002), and xieyi (P02003)), while some others are mutually shared by Chinese and Western cultures (e.g., oil painting (P02008)).

Secondly, the study conducted the analysis of alignment of NPM Vocabularies and AAT. The focuses were on: (a) whether AAT has provided a specific category that matches what is in the NPM Vocabularies; (b) whether a matching category in AAT has covered the values listed in a NPL unit completely.

For instance, the unit “Archaeological Culture (A02)” in the NPM Vocabularies can be mapped to the guide term <Chinese Neolithic Periods> in AAT, and the terms under both mostly have the exact equivalence relationship. However in many other cases, this ideal situation could not be found, as indicated in the patterns number 2 to 4 in Figure 1 below. Therefore this study had to combine different modules in the processes according to different mapping conditions in order to achieve the best mapping quality.

Findings

The study summarized the patterns of similarities and differences between the conceptual structures of the NPM Vocabularies and AAT, and amended the state of completeness of these aligned terms with each of the identified patterns.

When aligning the NPM values with AAT concepts, the study observed four patterns of similarities and differences between their conceptual structures (Figure 1):
Fig. 1: Patterns of conceptual structures for the Chinese art concepts covered in AAT

<table>
<thead>
<tr>
<th>Structures</th>
<th>Specific category in AAT</th>
<th>No specific category in AAT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Values</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Completely covered</td>
<td>Pattern#1</td>
<td></td>
</tr>
<tr>
<td>in AAT</td>
<td>A02 考古學文化 (13 values)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Archaeological Culture</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A17 時代</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chinese Periods (66 values)</td>
<td></td>
</tr>
<tr>
<td>Incompletely</td>
<td>Pattern#2</td>
<td>Pattern#3</td>
</tr>
<tr>
<td>covered in AAT</td>
<td>A03 窯系</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chinese Ceramic Style</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(19 values)</td>
<td></td>
</tr>
<tr>
<td>Not covered in</td>
<td>Pattern#4</td>
<td></td>
</tr>
<tr>
<td>AAT</td>
<td>P17 傳說動物</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fabulous Animals (7 values)</td>
<td></td>
</tr>
</tbody>
</table>

Pattern #1: there is a specific category in AAT; the values in a unit of the NPM are completely covered by that AAT category (e.g., NPM value vocabularies for A02 Archaeological Culture and A17 Period). For instance, the value Qing (Chinese style) is categorized under Chinese dynastic styles and periods and Yangshao is under Chinese Neolithic periods. This pattern may indicate that values represented by such thematic vocabularies are shared by the East and the West or that certain themes that are specific to the Chinese culture have already been accommodated in AAT. Pattern #1 is, therefore, the ideal pattern in which the concept and structure of two different structured vocabulary systems are highly interoperable. However, as also observed in the study, it is worth noting that certain correlations in terms of semantics, development or inheritance can be found by tracing notes of different vocabularies but not yet officially established in AAT. For example, relations with respect to the inheritance can be identified in the notes of Yangshao (ID: 300173481) and Longshan (ID: 300018347), but AAT has not established any kind of link between the two concepts to indicate such correlations. Nevertheless, we found that AAT has already created links between other similarly related concepts. For example, the relationship between First Dynasty (Egyptian) and Second Dynasty (Egyptian) are clearly identified by preceded and followed; the former preceded the latter and the latter followed the former. In addition, it is also noted that NPM tends to structure vocabularies in a flat, non-hierarchical way when compared with AAT; for instance, all values under Chinese Neolithic periods in the NPM belong to the same superordinate while those in AAT are structured in three hierarchical levels with hierarchical relations between them clearly indicated.

Pattern #2: although there is a specific category in AAT, the values in a unit in NPM are not completely covered by that AAT category (e.g., A03 Kiln Systems). This indicates some discrepancies. For example, NPM’s A03 Kiln System corresponds to AAT guide term Chinese ceramics styles. However of the 19 values under A03, several are not included in AAT (See Figure 2 below). The study further conducted a comparison between values specifically categorized under Chinese ceramics style by AAT and values in the unit of A03 Kiln Systems of NPM to explore differences between them. We found that in the
The differences consist mainly of values associated with Chinese exported porcelain, which contribute significantly in the history of Western art. Those concepts form an essential source for the West to approach and study Chinese porcelain. Nevertheless, they are not collected by NPM which features imperial collection of fine porcelain and establishes its system of vocabularies exclusively on the basis of its collection (Jörg, 1989).

Pattern #3: there is no specific category in AAT, and the values in a unit of NPM vocabularies are spread in various places in AAT (e.g., "A08 Chinese Ceramic Glaze by color"). It is clear that for specialized terms in the domain of Chinese culture and art, there are great discrepancies between the values in NPM vocabularies and AAT. Such a pattern is also found for the values associated with "A12 Technique." Pattern #4: there is no specific category in AAT, and the values in a unit of NPM Vocabularies found no equivalent concept in AAT. For example, "P17 Fabulous Animals" lists well-known legendary creatures found in folklores, literature, and the arts being unique to Chinese culture. There is neither a specific category for them or any concept covered in any category in AAT (See Figure 4 below).

To response to the four patterns of similarities and differences between these two conceptual structures, the study proposed possible solutions as follow: (1) adopting the conceptual structure of AAT as the primary structure and using that of NPM as a subsidiary structure; (2) adopting the conceptual structure of NPM as the primary structure and using that of AAT as a subsidiary structure; (3) adopting only the conceptual structure of AAT; (4) adopting only the conceptual structure of NPM.
Fig. 3: Situations where no specific category in AAT for the Chinese art concepts.

Fig. 4: Situations where there is no specific category and equivalent concept in AAT for the Chinese art concepts.

- Materials Facet
  - Materials (Hierarchy Name)
    - Materials (matter)
      - Coating (material)
        - Coating by location or context
          - Glaze
            - Ceramic glaze
              - Ceramic glaze by composition or origin
                - Blue and white (ceramic glaze)
              - Celadon (glaze)
            - Ceramic glaze by form
              - Clear glaze
            - Ceramic glaze by technique
              - Sang de boeuf

- Styles and Periods Facet
  - Styles and periods by region
    - Asian
      - East Asian
        - Chinese
          - Chinese styles (style)
            - Chinese ceramics styles
          - Chinese export
            - Famille (keramiekstijl)
            - Famille rose
            - Famille verte
          - Famille jaune
          - Famille noir
          - Wucai
          - Ding (Chinese ceramics style)
          - Ru
          - Shufu
          - Xing

- NPM (klin)
  - Ceramic glaze by color
    - Single color
      - Celadon (glaze)
    - Transparent glazed
    - Red glazed
    - Underslaze enamels
      - Underlaze enamels
    - Blue and white
      - Famille verte
      - Famille jaune
      - Famille noir
      - Famille rose
      - Wucai
      - Ding (Chinese ceramics style)
      - Ru
      - Shufu
      - Xing

- NPM (kiln)
  - Ding kiln
    - Famille verte
    - Famille jaune
    - Famille noir
    - Famille rose
    - Wucai
    - Ding (Chinese ceramics style)
    - Ru
    - Shufu
    - Xing

- Generic terms for the fabulous animals
  - Phoenix (phoenix)
  - Qilin (qilin)
  - Unicorn (unicorn)

- Generic terms for entities
  - Agents
    - People
      - People (agents)
        - People by occupation
          - People in the humanities
          - People in religion
          - People in religious occupations
          - People in animals
          - People in plants

- Generic terms for actual animals
  - Agents
    - Living Organisms
      - Animals
        - Mammals
        - Birds
        - Fishes
        - Vertebrates
        - Chordates
        - Puma (genus)

- Generic terms for actual animals
  - Agents
    - Living Organisms
      - Humans
      - Domestic animals
      - Wild animals
      - Fishes
      - Birds
      - Mammals
      - Chordates
Conclusion

The NPM Vocabularies and AAT were established under the principles of literary warrant and user warrant, with clearly different objects as the focus. With the efforts of the Chinese AAT-Taiwan team, the vocabularies are, for the first time, put together side-by-side. While reviewing published literature and the international standards on thesaurus construction and mapping, the authors encountered various questions that are beyond the processes of individual concept-to-concept mapping. This paper reported issues regarding the alignment of the controlled vocabularies in the domain of Chinese art as well as the patterns found in the effort of achieving semantic interoperability. The discrepancies in the conceptual structures for many concepts that are unique to the Chinese culture are significant as reflected in the NPM Vocabularies and AAT. Such a finding calls for attentions to those issues being different from concept-concept mapping, term-term translation, and string-string conversion. The findings in the paper are meaningful to the research on the semantic interoperability of multilingual KOS, especially when dealing with cultural-related concepts that cannot be exactly aligned in vocabularies due to the discrepancies in the conceptual structures.

References


Discourse Analysis as an Approach to Categorizing the Domain of Public Policy:
The Case of Brazilian E-Government

Abstract
Taking the discourse as our empirical object of study, we used the theoretical foundations of the Critical Discourse Analysis (CDA) to justify the way an ontology is built to represent the area identified as a public policy, in this case, the Brazilian Electronic Government Program. Our objective is to contribute to the fields of Information Science and Computer Science, and present the possibilities for knowledge representation and organization of knowledge by Ontology construction. To support the construction of an ontology for Brazilian E-Government, as a public policy, we evaluate the proposed Controlled Vocabulary of E-Government - VCGE (2011).

Background
The way to represent the area identified as public policy through ontology is still an experiment in the field of information science. Public policy can be generally defined as a system of laws, regulatory measures, courses of action, and funding priorities concerning a given topic promulgated by a governmental entity or its representatives. However, we believe that public policy is a process resulting from conflicts between social groups, interests and discourse.

There is consensus that ontologies serve as a model of knowledge representation in the fields of Information Science and Computer Science, and present themselves as appropriate tools for knowledge representation and organization of knowledge in information systems as they seek to organize and standardize concepts, terms and definitions accepted by a particular community.

Soergel (2004) highlights the lack of communication between communities that deal with conceptual structures, where the efforts for building products are fragmented, resulting in considerable reinvention of concepts. Silva (2008) demonstrated problems related to lack of a standard for the construction of ontologies in addition to lack of systematic explanations of how, where and under what limits theoretical approaches may be used within the development process. In the context of building controlled vocabularies, the theoretical principles that govern their constructions are not explicit (Silva 2008). The principles stated in the theory of classification (Ranganathan, 1967) and the theory of concept (Dahlberg, 1978) are useful. The development of an ontology demands political, epistemological and conceptual decisions (Campos et al., 2007). When building an ontology one tries to reach an agreement on the use of shared vocabulary in a coherent and consistent manner.

The most widely quoted definition of ontology within knowledge representation is Gruber's (1993b), in which the author states that an ontology is a formal and explicit specification of a shared conceptualization, which brings us to the act of communicating, as communication is sharing of meaning. To support the construction of an Ontology for Electronic Government as a public policy, we evaluate the proposed Brazilian Electronic Government Program and some tools already available such as the e-PMG - Electronic Government Metadata Standard and the e-PING - Electronic Government Interoperability Standard, notably the Controlled Vocabulary of E-Government - VCGE (2011).
The VCGE was built as a scheme to be used in the subject element/category of the Electronic Government Metadata Standard (e-PMG) and proposed to: define a conceptual or ontological framework; identify and select the significant information of each document; advise users in identifying their information needs, facilitating and enriching their search. It was noted, however, that the conceptual system on which VCGE's construction relied, despite being focused on the citizens, chose public administration to be represented and, because it is a broad concept, it is not adequate to incorporate the "Discourse" of the Brazilian Electronic Government specifically.

Moreover, the conceptual system that defined the VCGE does not take as its reference a conceptual network like the conceptual structure (infra and supra) of a domain of knowledge and the theoretical discussion that identifies information as a relationship between two places, or states: Center and periphery (Latour, 2006).

**Objectives**

This article intends to reflect on a proposal for the representation of public policy and for the organization of knowledge based on the evaluation of Controlled Vocabulary VCGE Electronic-Government (2011), prepared at the heart of E-Government Brazilian Policy with the objectives of, firstly, to replace the list of Government Affairs (LAG), and in second place to operationalize the Government Metadata Standard (e-PMG), an instrument to implement this policy Interoperability Standard. The conceptual network proposed here is based on an analysis of the discourse found in the Brazilian Electronic Government Program and the main concepts identified making use of such categories as information policy, and information as a relationship of domination, subordination and exchange.

**Context**

Since 1995, when the Internet in Brazil was structured and grew beyond the restricted academic realm, the movement was seen as a way for the Brazilian government to use the new information technologies in service delivery and provision of information to citizens, providers and servers, as a move towards e-government.

As Jardim (2000), "e-government strategy expressed by the state apparatus makes use of new technologies to offer the society better access to information and government services, increasing the quality of services and ensuring greater opportunities for social participation in the democratic process”.

According to Gontijo (2002) "The Brazilian project of e-government is emerging as an information policy and is subject to analysis from the concept of Information Scheme. In Brazil, the e-government is attempting to stabilize conflicts between social groups and interests. It is legitimate, and the policy of information is there to aim to change the current informational context characterized by the digital divide. Two approaches are emphasized and, although not contradictory, the predominance of instrumental approach may result in the preservation of the digital divide of the country”.

In the definition given by Gonzalez Gomes (1999), an Information Policy, in its broadest form, can be defined as the set of actions and decisions aimed to preserve and reproduce, replace or change an information system, and can be either tacit or explicit policies, micro or macro, and in principle the focus of its manifestation would be the state and public policies.

**Methodological Procedures**

The area of Knowledge of Representation and Organization involves two fundamental concepts: the organization of knowledge and representation of knowledge. Its object of research is knowledge, and its main concerns around that object, are organization and representation. The representation of knowledge is seen by Dahlberg as the logical structure
of the conceptual representation, and also the result of the identification of concepts that we
determined on the basis of the terminology used.
Dahlberg (2006) believes that knowledge is subjective, individual in nature and it is not
transferable and may only be prepared by a personal reflection of someone declaring
that knowledge “and may only be acquired through reflection”.
Critical Discourse Analysis is a practice and a field of linguistics and of communication
focusing on analyzing ideological constructions found in a text. Every discourse is a social
construction - not an individual one - that can be only analyzed considering its context, its
conditions of production; it also means that discourse reflects a particular world view,
necessarily linked to its author(s) and to the society where they live in. The empirical object
of discourse analysis is the construction in which the analyst works, in order to find, on its
surface, the signs that guide scientific research. It should be noted that the object of
Discourse Analysis is Discourse.
The interpretation of a discourse should know, immediately, that there is an author, a
subject with a particular social and historical identity, and based on this fact, place the
discourse as sharing this identity. It should be highlighted that the text will only receive this
nomenclature/label (text) if the recipient of the message can decode it.
From this perspective, we examined Critical Discourse Analysis (CDA) which is an
interdisciplinary approach for studying texts, that views "language as a form of social
practice" (Fairclough 1989) and intends "to reveal the ideological foundations of discourse"
(Teo 2000). One of the ideological foundations to be unveiled by the Brazilian Electronic
Government Program discourse is what Latour (2006) identifies as making information
available, which takes into account a relation of domination, of subordination. The author
conceptualizes information as a relationship between two places, or stages: center and
periphery (Latour, 2006).
Also according to the author, it is impossible to understand information from this
perspective without taking into consideration the institutions that allow the establishment of
such relations and without the material vehicles (inscriptions) that transport and load them.
The production of information allows the solution of the contradiction between the
presence and absence of a place, in a practical way, by operations of selection, extraction
and reduction. These operations allow, thus, the reduction of a certain reality and then its
expansion afterwards, through the representation of issues hitherto isolated on a single
plane. Such operations are focused on what he calls as the calculation centre, explains
Another ideological foundation highlights the instrumental nature of information policies
that restrict themselves to solving problems such as technology implementation, improved
communication between departments within the public administration, increased access to
governmental documents, rather than policies that focus on the relationship of information
and power, or how power is exercised in and through social relations mediated by
information (Frohman 1995).

Case Study
In this perspective, the concept of electronic government, due to its eminently political as
well as public nature, must necessarily be perceived as a process in constant development,
promoting institutional changes and government transparency which, at the same time,
reinvent the real government itself.
E-government could be considered within a broader perspective, as a possible key to
promoting accessibility to key information for the articulation of support, training of many
pressure groups, civic capacity and social capital increase, in addition to promoting
economic development and more democratic and transparent relations between government
and civil society. Overall, government relations grow with the other actors in both virtual
and real terms (Rüdiger, 2002).
However, when analyzing the controlled set of terms the descriptors of the Controlled Vocabulary for Electronic Government - VCGE show us the following representation of the related program:

<table>
<thead>
<tr>
<th>Label</th>
<th>VCGE – First Label</th>
<th>VCGE – First and Second Labels</th>
<th>VCGE – Third Label</th>
</tr>
</thead>
</table>
This can be represented as:

**Fig. 1: Focus on Public Administration**

![Diagram](image1)

**Conclusion**
We examined the concept of a system of information as mapping a process resulting from conflicts between social groups, interests and discourses in order to describe the Brazilian proposal of Electronic Government (E-Government) as an information technology and communication policy which is public administration oriented.

The proposed conceptual network is organized along the following guidelines:

1. Promotion of citizenship
2. Digital Inclusion is inseparable from the E-Government
3. Free software opens the field of production and circulation of knowledge
4. Knowledge management to ensure and share strategic knowledge
5. Rationalizing the use of resources
6. Integrated framework of policies, systems, standards and norms
7. Integration shares power

If the focus in the E-Gov program are the guidelines above, this formatting can be represented as:

**Fig. 2: Focus on Information Policy**

![Diagram](image2)
References
The CPDOC Semantic Portal: Applying Semantic and Knowledge Organization Systems to the Brazilian Contemporary History Domain

Abstract
Presents the semantic portal project of the center for teaching and research in the Social Sciences and Contemporary History (CPDOC) of the Fundação Getúlio Vargas, Rio de Janeiro. This project involves the use of semantic and visualization technologies and natural language processing techniques to allow enhanced ways to access the CPDOC collections.

Introduction
CPDOC is a major center for teaching and research in the Social Sciences and contemporary History located in Rio de Janeiro, Brazil. CPDOC is also the leading research institute in the country and holds a major collection of personal archives, oral histories and audiovisual sources pertaining to Brazilian contemporary history. It is part of Getúlio Vargas Foundation (FGV), a prestigious Brazilian research and higher education institution founded in 1944.

This paper describes the semantic portal project being developed at CPDOC and its partner, the Applied Mathematics School from FGV, along with all the initiatives that are being undertaken in order to achieve the final goal. Among those initiatives we can highlight the development of two ontologies; one for the field of Brazil’s contemporary history and another one that aims to standardize all document metadata throughout the CPDOC archives. The final goal is to implement the descriptive information about the archives in a RDF triplestore, applying semantic technologies, and connecting the digitized archives to the Linked Open Data (LOD) hub, granting access to the World Wide Web users via a single and unified interface. Inspiration for this initiative was the STAR project (Binding et al, 2010), that used similar techniques in the archeology domain.

The Archives
The CPDOC archives are multimedia in nature, being composed by many collections comprising journals, books and documents in audio, text, video, pictures, drawings and photographs. The main collections are organized as such:

- Personal Archives (AP): About 200 archival files, summing up to 1.8 million documents including text, images and videos.
- Oral History Program (PHO): A huge set of testimonies (in audio and video) consisting of more than 1000 interviews, which add up to five thousand hours of recordings
- Brazilian Historical Biographic Dictionary (DHBB): the current version includes 7553 entries, of which 6584 are of biographical nature and 969 relate to institutions, events and concepts of interest in Brazilian history after 1930.
In 2008, the CPDOC initiated a huge digitization project, converting most of the archives to electronic format. In 2010 the archives had 300,000 textual documents, 65 film rolls, 106 tapes (VHS, Beta and U-MATIC), 350 Long Plays, 187 K7 tapes, 85 roll tapes and about 32000 photographs from the personal archives files. Besides that, 5000 hours of interviews from the oral history program were also digitized and all these documents are available for consultation by researchers locally in the institution. When the digitization is completed approximately 80,000 photographs will be available for Web access, comprising the whole set of images donated to the Center. Additionally, all the DHBB are already in the digital format.

The problem
The main problem to be addressed is the heterogeneity and multimodality of the archives, and the lack of descriptive metadata standards for all types of document. For a semantic portal to work, we do need a single and unified standard to describe each item in the archives, and also a domain ontology - in this case, the Brazilian contemporary history. The three main collections, or „systems”, as they are called internally – AP, PHO and DHBB – have different sets of controlled vocabularies of their own. As they were built independently over time and by various different groups of researchers and analysts, each one of the three presents a great degree of semantic idiosyncrasy. The lack of links between concepts – internally to the systems, not to mention external relationships with other collections outside FGV – undermines the exploitation of the myriad of relations among the items, making every collection insulated. Besides that, and in spite of the single search interface it shows today on the website, the technologies used are independent too, impeding cross search among the collections. The main search tool is a full text index, that yields high recall but low precision, and the net effect is that the collections are underutilized.

The proposed solution
The presented scenario is a complex one, and requires a number of reasonably independent initiatives. In the following sections, we will present the solution's goals, the main set of tasks to be fulfilled and the initiatives taken so far. The main goals are:

- To offer unified and integrated access to the whole set of CPDOC systems (AP, PHO and DHBB);
- To give the users a rich set of navigational options: concept or hierarchy based navigation, and also full-text search;
- To enable Searches that traverse different systems, finding all relevant results – achieving good balance between precision and recall;
- To be able to present the search results in a faceted manner by the most relevant metadata elements (e.g. media format, collection name, etc)
- Homogenize concept terminology – no more syntactic idiosyncrasies;
- Adopt widely used metadata formats and ontologies (e.g. DCMI, FOAF, Bibliographic Ontology, COMM, EAD);

1 http://cpdoc.fgv.br/acervo/arquivospessoais
2 http://cpdoc.fgv.br/acervo/historiaoral
3 http://cpdoc.fgv.br/acervo/dhbb
4 http://dublincore.org/
5 http://www.foaf-project.org/
6 http://bibliontology.com/
7 http://comm.semanticweb.org/
8 http://www.loc.gov/ead/
• Development of specific ontologies for the domain of the Brazilian contemporary history and enhance documents description by the addition of uncovered characteristics / attributes;
• Building a RDF triples database – each triple will connect a document and its metadata, allowing integration with other cultural heritage resources via the Linked Open Data (LOD);
• Migration of the current knowledge organization systems (KOS), as glossaries, thesaurus and controlled vocabularies to the SKOS format and integration to the ontologies. All resources will have an URI;
• Promote enhanced visibility to search engines, through revamping the portal interface and eliminating the current need to login to access;
• Facilitate the process of document tagging with the new metadata standards; as there will be a huge amount of documents that will need to be analyzed and described;
• Integrating with the Learning Objects Databases and the FGV Digital Library.

These goals are the main guidelines to actions and to realize these there are several ongoing projects and initiatives. These initiatives are described below:
1. Face and character recognition project: Has as its main goal to offer an optimized process for tagging the images in the AP archive, by means of computational facial recognition of characters. A software was developed to apply facial recognition techniques to associate faces with characters described in the image caption – an idea pioneered by Google's Picasa software. This software allows for a friendly way to browse the AP photography archives, together with the identification of the characters in each photograph. As the faces are recognized, each photograph would generate a RDF triple linking the photograph URI to the KOS that contains the names of relevant personalities in Brazilian contemporary history. The main interface of the developed solution is shown in Fig. 1.

Fig. 1: The VIF (Very Important Faces) software main window – character's tagging

9 http://linkeddata.org/
10 http://www.w3.org/2004/02/skos/
11 http://picasa.google.com/
2. Sound and text alignment project: Its goal was to produce automatic alignments and, therefore, transcriptions of voice and video recordings for the Portuguese language, in the scope of the PHO archives. The recorded audio consists of interviews. Some of the recordings have already been transcribed by humans while others are available as audio only. This process would link the excerpts of interviews to the transcriptions, and these transcriptions would be harvested using natural language processing techniques to extract common concepts – allowing it to be linked to the existing ontologies.

3. Text mining project: Consists of a set of initiatives conceived to give support to both the “Face and character recognition” and “Sound and text alignment project”. Here, important descriptors are mined from image captions, associated documents, and transcriptions, thus generating RDF triples associating to the domain ontologies.

4. DHBB “Wikification” project: Was conceived to promote a deeper interconnection of CPDOC’s internal databases with external ones, such as Wikipedia. The benefits of this interconnection are improvements in public visibility and the establishment of social collaborative networks to contribute/correct the collections. It is being implemented with the open source semantic wiki software MediaWiki 12 with its semantic extensions 13. Entries are being migrated from the current system to the wiki to demonstrate the capabilities of the functionalities of the tool. This project benefits from the ontologies that are being created, that will serve as semantic annotations to the entries.

5. Contemporary History Ontologies: after realizing the lack of formal KOS resources regarding the domain, it is being developed as a master’s dissertation, the definition of a conceptual model for contemporary history ontologies, covering competency questions that all ontologies must deal with, and the main facets: Agent (person, group or organization), Role, Event, Place, Time and Document (source of historical information). The huge (10,000+), non-hierarchical list of concepts currently used for the description of the archives, along with the entries of the DHBB, will be good sources for the creation of domain ontologies. The first light ontology developed under this model deals with the “military government years”, a phase in Brazilian history also called as the “dictatorship years”.

6. Description Ontology: Another ontology, the CPDOC description ontology is being developed as part of a doctoral research. This research includes analysis of the descriptive metadata used in each archival file nowadays; the harvesting of existing metadata standards for all kinds of documents and media (Dublin Core, Bibliographic Ontology, FOAF, COMM), and issues of compatibility with archival description standards (NOBRADE, ISAAR(CPF), ISAD(G)).

7. Implementation of a RDF triplestore: in order to store the triples built – all the triples associate a document URI with an ontology concept URI. There will be set a Sesame 14 database to serve as the triplestore. We will have to generate URIs for each document item stored in the current Oracle database that holds all the digitalized documents.

All the projects are independent, in spite of the fact that they converge and aid the idea of a Semantic Portal, as depicted in Fig.2.

The main activities described earlier are shown in the framework above. In the leftmost we have the CPDOC archives, that is being digitized and, in this process, natural language processing techniques are being used to find all the access points they will share, and the

12 http://www.mediawiki.org/wiki/MediaWiki
13 http://semantic-mediawiki.org/
14 http://www.openrdf.org/
concepts for linking are brought from the domain ontologies that are being developed. The RDF triplestore connects those ontologies to each item in the collections, and allows the portal interface to offer concept based and hierarchical search, along with a SPAQRL query builder for advanced users. The portal architecture is still being discussed.

**Fig. 2: The Semantic Portal Framework**

**Current Developments**

The whole project began in 2010, with character recognition tasks. Nowadays, many tasks are being developed in parallel. The researchers’ academic backgrounds involve mathematics, history, engineering, information science, computer science archival science, librarianship and even movie science and biology. Those are present in the main groups, with the tasks of developing KOS (ontologies and thesaurus), NLP tasks, image processing, sound processing and automatic RDF triples creation. In spite of some budgetary constraints, there are good results so far, and we do expect that the portal can be fully operational by the end of 2013. As independent projects, each one will offer functionalities and features that can enhance access to the archives. We hope that the project can offer a new resource for a huge community of researchers in the field, and also add one of the first Brazilian contributions to the Linked Open Data information hub.

**References**

Challenges of Knowledge Representation in Contemporary Archival Science

Abstract
Since its emergence as a discipline, in the nineteenth century (1889), the theory and practice of Archival Science have focused on the arrangement and description of archival materials as complementary and inseparable nuclear processes that aim to classify, to order, to describe and to give access to records. These processes have their specific goals sharing one in common: the representation of archival knowledge. In the late 1980 a paradigm shift was announced in Archival Science, especially after the appearance of the new forms of document production and information technologies. The discipline was then invited to rethink its theoretical and methodological bases founded in the nineteenth century so it could handle the contemporary archival knowledge production, organization and representation. In this sense, the present paper aims to discuss, under a theoretical perspective, the archival representation, more specifically the archival description facing these changes and proposals, in order to illustrate the challenges faced by Contemporary Archival Science in a new context of production, organization and representation of archival knowledge.

Introduction
Knowledge Organization (KO) as a field that studies the processes and systems of document description, indexing and classification finds in Information Science (IS) a fertile ground for discussions on “nature and quality of such knowledge organizing processes (KOP) as well as the knowledge organizing systems (KOS) used to organize documents, document representations, works and concepts” (Hjørland, 2008).

This paper aims to discuss, from a theoretical perspective, archival representation, more specifically the archival description in order to illustrate the challenges faced by Contemporary Archival Science in the new context of production, organization and representation of archival knowledge.

Since its emergence as a discipline, in the 19th Century (with the publication of the Manual for the arrangement and description of Archives in 1889 by the Dutch trio Muller, Feith and Fruin), the theory and practice of Archival Science have focused on the arrangement and description of archival materials as complementary and inseparable nuclear processes that aim to classify, order, describe and provide access to records. These processes have their common goal: the representation of archival knowledge.

Over the years, the archivist was seen as a simple guardian of records and documents, whose role on knowledge arrangement and description should be limited to revealing the meaning and the significance explicit in the document. In this sense, the archival knowledge representation was also restricted to the immediate and static meaning of the record, to the things one could read directly in the document.

In the late 1980’s there was a paradigm shift in Archival Science after the appearance of new forms of record production and information technologies. There was rethinking on its theoretical and methodological bases founded in the nineteenth century. Postmodern Archival Science was born, aiming to “denaturalize” what the modern Archival Science hitherto assumed to be natural, normal and rational. The core definitions and processes of Archival Science were put into a discussion and brought to debate the concept of archival record and the processes used to create, organize and represent it. The archival record, once understood as a sub product of an innocent, neutral and impartial action – features proposed by Sir Hilary Jenkinson in 1922 – was now seen as an entity socially and culturally constructed, as a symbol shaped by an author for a specific purpose, far from
being a sub product of an innocent and impartial action. Consequently the organization and representation of this knowledge should no longer be studied as neutral and free of functional and social influences.

Specifically regarding the representation of this knowledge it is important to highlight the role of the archivist as an active agent in memory construction and meaning attribution, since the record meanings are now being constantly reviewed as each user uses it to different purposes. From the Postmodern Archival Science perspective, the record meaning changes over time according to the uses to which each user intends to do from it. As stated by Duffy and Harris (2002, p. 265) “records are always in the process of being made, "their" stories are never ending, and the stories of those who are conventionally called records creators, records managers, archivists, users and so on are (shifting, intermingling) parts of bigger stories understandable only in the ever-changing broader contexts of society”. In this sense, the Archival Science moves its focus from the set of records to its social, organizational and functional on-going creation contexts.

As a result, the act of representation must follow a dynamic flow and archivists should “begin to think less in terms of a single definitive, static arrangement and description process, but rather in terms of continuous, relative, fluid arrangements and descriptions as on-going representational process” (Yakel, 2003).

The principle of respect des fonds and its challenges in the Post-modern archival representation

The principle of provenance (respect des fonds) since its promulgation in 1841, in order to solve the problems caused by the subject organization adopted by the Archives Nationales in France after the French Revolution, is considered to be the guide to organizing and representing process in Archival Science and perhaps the safest method to preserve the integrity of the sets of records produced by the same person or institution. It consists to “group, without mixing them with others, the archives (documents of every kind) created by or coming from an administration, establishment, person, or corporate body” (Duchein, 1983). Though its importance and precision have never been contested it is important to understand that, facing new perspectives of creating, organizing and representing processes, the principle also goes through changes that should be reflected in archival representation.

As pointed out by Duffy and Harris (2002, p. 268) “in modern bureaucracies, it is common for the same records to be created, accumulated, and used by numerous, different, successor or parallel agencies. Records emanate from business activities and in turn are used to support and carry out other business activities. Moreover, series of records move from the control or custody of one organization to another. This reality has led numerous archivists to suggest that the multi-faceted aspects of provenance are eroded when archival practice dictates the creation of fonds-level description and credits the creation of the records (and thus provenance) to one, and only one, individual or organization”. There is therefore a need to connect the record to all its contexts, focusing on all the relations that were established between it and the entities that use it, so that its multi-faceted nature can be represented entirely.

In this sense, the provenance study as an assumption to archival representation finds a fulcrum not only in understanding the record creator, but also in the relationship between the creators, the functions and management and maintenance systems of records. Once again, all the contexts should be represented at the moment of classification, arrangement and description of archival knowledge.

In the classification system usually adopted by the Archival Science the fonds (also known in Anglo-Saxon countries as “record group” or “archive group”) are the broadest and most essential units of record to Archival Science, followed by the records series and
the item. However, over the last decades the application of the record group concept for archival description faced a practical problem that derives, as pointed out by Cook (1993, p. 24) “from viewing it exclusively as a physical entity rather than as a conceptual principle”. This view reflects a records-oriented, descriptive cataloging tradition, as opposed to the context-oriented life-cycle data management approach. In other words it is essential that the archivist can focus on the context rather than on the record, understanding the principle of provenance as a virtual and dynamic principle that rules the practical activities and the establishment of physical series. In Bearman and Cook’s view it is essential that archival description focus on “the conjunction of the context of the activity and the information system in the records creating organization. This conjunction might take place at the level of the fond” as long as the fond is truly presented as a conjunction of the creator’s functions and activities on the one hand and, on the other hand, of the records and information systems - the actual products - which proceed from those functions and activities” (Cook, 1993, p 28) - or at the level of the series.

The postmodern approach and its authors defend the abolition of the record group from the archival classification system, once it enables the archivist to work with the multi-provenance nature of records. As they understand the record as a continuum, to establish a one and only creator or fond would be too limiting. To McKemmish (1998, p. 192) ”the physical reconstruction of the fonds in a record group, while providing one view of what is a multiple reality, obscures or obliterates other views”. In order to understand the concept of fond as a virtual and conceptual guide – instead of a physical category - some authors have proposed the series to be the first category for purposes of arrangement in archives.

This approach based on functional-structural contexts proposed by postmodern Archival Science finds a place, among others, in the Australian description systems that have been investigating these contextual relationships understanding the archival records as dynamic and virtual organisms that are constantly evolving over the last decades. The description system of the Australian National Archives since the publication of Scott’s articles about the abandonment of the record group concept, in 1866, ignores the record group concept describing records series as the primary level of classification and the item as the secondary level. According to Scott there are some problems presented in the application of the record group concept to archives (i.e. when the record group is created from the records of a transferring or controlling agency the risk is that they might lose their creation original order; or when similar series are created by different agencies in the same organization) that can only be solved with the application of the series as a primary category.

As regards to archival representation considering the series as a primary physical category in the classification systems allows the archivist to

quote

“describe records series in their totality and links descriptions of records to all the contextual entities that created, accumulated, used, controlled, owned, or transferred the records in the series. This system emphasizes the importance of linking a record entity to its various contextual entities and stresses the importance of inter-relationships” (Duff; Harris, 2002, p. 268).

Some authors like Eastwood (2002) and MacNeil (1992) do not agree with the abandonment of the record group concept advocating that there is no other provenance beyond the agency that creates the record. For these authors the concept of fond or record group fits perfectly.

footnote

1 According to the Glossary of Archival and Records terminology, published by the Association of American Archivists, the series are groups of similar records that are arranged to filling system and that are related as the result of being created, received, or used in the same activity.
Although the series system (postmodern) and fonds-based approaches disagree regarding the provenance issue, both pursue the same goal, which is to represent – organize and describe – the context and the evidential value of the archival records.

The power to describe is the power to make and remake records and to determine how they will be used and remade in the future. Each story we tell about our records, each description we compile, changes the meaning of the records and re-creates them. These different views of provenance significantly affect the type of descriptive architecture proposed by their advocates.

Equally influential are their assumptions about what archival description is, when description takes place, and its purpose (Duff; Harris, 2002, p. 271).

The connection between the record and its context is an intellectual process that is materialized through the finding aids resulting from the archival knowledge representing process and understood as a bridge between the user and the record. According to Cook, one of the postmodern authors the description must reflect the ever changing records history, since each time the user accesses the record new representations are created from new uses and interpretations. In this sense the finding aids must be constantly recreated and updated to reflect and to represent complex and dynamic realities in ever changing process. On the opposite side there is the fonds-based description that is based on bibliographical description, which represents static objects producing static finding aids.

The representing process in Archival Science is also based on international standards for archival description and encoding of finding aids such as ISAD (G) – General International Standard Archival Description – ISAAP (CPF) – International Standard Archival Authority Record for Corporate Bodies – and EAD – Encoded Archival Description – that present a challenge to information professional in a context where the desconstruction of the standardization of routines is necessary as the record production, organizing and representing processes acquire a dynamic status in a postmodern discipline, going against the static and neutral position observed so far.

Despite these international standards for archival description, the Archival Science faces a conceptual challenge, once its principles and definitions not always find a consensus between the authors in the area. The principle of provenance and the record group/fond concepts are only a few examples among lots of concepts and terms that are interpreted in many different ways. This conceptual problem has a mayor reflection on organizing and representing processes once a concept or principle miscomprehension can lead the professional of information to irreparable mistakes. That is why is so important and urgent to discuss these issues, once it allows the solidification of the discipline in a moment of paradigm rupture and changes, facing the new forms of records creation and management.

However the biggest challenge facing the Archival Science today regarding the knowledge representation is to make the archivist understand the complex system of relationships between the record and its historical context (i.e. all the existing intentions behind the record creation and the information it carries) so he can move “from a monolithic legacy of past archival theory, from the old fashioned „one-thing-one-entry approach” (Cook, 2001, p. 32).

Conclusion

Considering these questions, it is possible to conclude that the postmodern approach is the only one that seems to deal better with contemporary issues, once it goes deeper on questions about archival representation and its relationship between the archivist, the record creation context and the user. In this sense, the archival postmodern representation “would reflect, in short, sustained contextual research by the archivist into the history of the records and their creator(s), and produce ever-changing descriptions as the records creation and custodial history itself never ends (as at the moment of archival accessioning or of creating
The archival knowledge representation can no longer be limited to standards of archival description or static views proposed by the archivist and the record itself. The contemporary Archival Science must be aware that archival representation is continually reinvented, reconstructed and reborn.

To describe is how the archivist make and remake the record. In this context it must be his role to find between these two or more approaches a balance that can offer the best record representation, one that can fit in the very particular and unique archival context which he/she is describing.

References
Duchein M. 1983, Theoretical principles and practical problems of respect des fonds in Archival Science, Archivaria, 16 (summer): 64-82.
MacNeil H., 1992. The context is all: describing a fonds and its parts in accordance with the rules for archival description, In: The archival fonds: from theory to practice, ed. Terry Eastwood (Ottawa: Bureau of Canadian Archivists).
Archival Classification and Knowledge Organization: 
Theoretical Possibilities for the Archival Field

Abstract
The main goal of this study is to outline a possible relation between archival classification and knowledge organization theory. In this sense, we seek to contribute to the conceptual classification in Archival Science, since there is a lack of systematization about archival classification; not just classification, but even the study of historical and conceptual aspects of the discipline. In the context of knowledge organization there is a considerable amount of research on how to build classification schemes and indexing systems that can help contribute to and expand archival classification theory. In order to comprehend this vast field of theories and methodologies we construct a parallel comparing the classification concepts in both areas and analyzing these concepts.

Contexts: The Archival Classification
The Archival Science has been looking to demarcate its actuation space, with studies related to a reconsideration of its methods and techniques. We seek in this article to demonstrate some possible contact points between archival classification and knowledge organization especially in terms of contributions from the second for the theoretical growth of archival classification. Thus this work is also related to conceptual classification, since there is a lack of systematization about archival classification. Theoretical discussions are essential for maturation of the discipline in order to fill gaps in historical and epistemological frames.

Several authors have written on classification as a concept and its theoretical and practical applications, e.g. Laura Millar (2002), Terry Cook (2004, 2005), Brien Brothman (1999, 2006) e Chris Hurley (1995a, 1995b, 1998, 2000) Tennis (2010, 2011) and Dahlberg (1993, 2006). In the context of knowledge organization there is a considerable amount of research trying to understand how to build classification schemes and indexing systems. These studies can, in a broad sense, help to contribute and expand archival classification theory. So we try to link the discussion of functional classification promulgated in the current archival theory, with the development of classification schemes and classification as an area of study in knowledge organization (see for example, the articles by Tennis (2010, 2011).

So, we seek to compare the notions/concepts of classification not to find a definition or a unique sense for both areas, but to understand the differences and similarities in thinking about a reconsideration of archival classification, but also having in mind the archival classification specificities. In archival field there is a plurality of terms to designate the activity of classification of documents.

Currently Archival Science – in terms of both theoretical and practical aspects - is undergoing profound transformation. A number of publications have tried to define the boundaries of the field and there is also research on functional classification for the organization of modern records. These have been studied in Australia and Canada as a response to the current understanding of the archival institutions. We can highlight the important contributions of Laura Millar (2002), Terry Cook (2004, 2005), Brien Brothman (1999, 2006) and Chris Hurley (1995a, 1995b, 1998, 2000), Terry Eastwood (2000) responsible for the changes in automated description, assessment and functional analysis of the records. All these changes have implications for understanding the concept of
provenance vis-à-vis *respect des fonds*, affecting the processes of classification, because provenance is the guiding principle of classification. Millar (2002, p.6) highlights this:

> The fonds implies a wholeness, a completeness, a totality. I would argue that no archives now has, ever will have, or ever has had, "the whole of the records" of any creating agency. Records are destroyed, or lost, or transferred, or changed even before they get to the archives. Once they are in custody, they may be culled, and weeded, and selected. Archivists don’t just manage records; they actively decide what will be kept and what will be removed, through the very process of appraisal.

This concept of provenance has caused changes in the understanding of the concept and process of classification; This new way of understanding is based mainly on a more elastic concept of provenance as we can see in the texts of Hurley (1995, p. 22), one of the major authors in this context:

> Context comprises both the data carried by the record and the knowledge brought to the record by the user. Contextual knowledge forges the link that is the basis of understanding. Efforts now being made to regularize the process whereby knowledge of context is captured as metadata for electronic record-keeping should not blind us to a fundamental truth. Because records themselves are time bound, metadata must be verified within a context that is both current and historical.

Therefore, the provenance is not understood narrowly as in the works discussed above, it can mean different things depending on the context in which documents are included, the institution and its users. Hurley (1995, p.10) complements this idea by talking about his understanding of the functions: "Functions themselves have a history and a character independent of the recordkeeping agent that is being described."

And in this sense, this is a major difference in the archival classification due to the materials of an archive.

**Contexts: Knowledge Organization and Archival Classification**

In the context of the current practice of classification, it is possible to say that archives organization resembles what is practiced in the knowledge organization (KO) field to the extent that knowledge organization is a conceptual-theoretical framework that deals with recorded and socialized knowledge. This supports the view that exists in KO, classification, concept theory and other areas of information retrieval with focus on representation of descriptive and subject knowledge irrespective of the form of the resource – physical, digital and virtual.

For Barité (2001, p. 38 our translation), the KO, while a professional activity, "[...] aims to present theoretical support (and feedback on itself) with all that is relative to information treatment, particularly with the treatment of thematic information, and less specific - but not less important - with the management of the social use of information."

Thus, the KO area, "[...] is used primarily for socialized knowledge [...]" (Barité, 2001, p. 41 our translation) and registered, thus played a crucial mediating role between a production context and use of information.

As a discipline, "[...] covers the development of techniques for construction, management, use and evolution of scientific classifications, taxonomies and documentary languages. On the other hand, brings methodologies, use and retrieval via natural language "(Barité, 2001, p. 41, our bold and translation).

In this sense, KO, as Guimarães (2000, p. 210 our translation) suggests, that the KO field works with circle "[...] considering the possibility of organizing a recorded knowledge from the perspective of generating new knowledge that, once registered, will transform in information (knowledge in action, in the design of Dahlberg, 1993, p. 214) to generate new knowledge."

The knowledge according to Dahlberg (2006, p.12) can be defined by its construction processes or its degrees of complexity:

- Knowledge elements, by which we understand the characteristics of concepts that can be gained by predicating the properties of or making statements about referents (characteristics as knowledge
elements – elements of knowledge units (concepts) – should not be confused with features of concepts, e.g. broader, narrower, related, etc.);

- Knowledge units, which we equate with concepts. They are the synthesis of the concept characteristics, gained by said statements about referents and represented by a sign (word, name, term, code);
- Larger knowledge units, which are concept combinations, e.g. in statements or in definitions or just in texts; and,
- Knowledge systems, which are entities composed of knowledge units arranged in an adequately planned, cohesive structure

Esteban Navarro and Garcia Marco (1995, p. 149, our translation) have a good definition for what we can related to the KO field:

[...] devoted to the study and development of fundamentals and techniques of planning, construction, management, use and evaluation of systems for describing, cataloging, classifying, ordination, storage, communication and retrieval of documents created by man to testify, preserve and transmit their knowledge and their actions, from its contents, in order to ensure its conversion into information capable of generating new knowledge. It is therefore a science-dimensional, since it focuses on principles, methods and instruments put into action for the management of human knowledge from a triple perspective: their representation, their organization and their communication as documents. Nevertheless, the Science of Representation, Communication and Knowledge Organization, called in a common sense Knowledge Organization, due to which the organization is a mediator between the other two acts, since, on one hand, representation is carried out in order to allow an effective organization, and, second, communication requires a proper recovery, whose success depends on the quality of the organization

As said in the in the quotations, it is possible to related the archival classification and its practice to the KO field, in - organization – as an fundamental intellectual activity, for the structuralization in classes of records and documents. We can see the classification as a key practice for the proper accesses and retrieval of documents and records, the archival science lacks of terminology and methodologies for the scientific study of archival classification, the KO area have a major frame work in that area and can help the study of classification in the archival field, in most prestigious way.

Terry Eastwood (2000, p.93-94 our bold) endorses that when he says:

The choice of the word “arrangement” as the name of this process is unfortunate. It denotes placing things in proper, desired, or convenient order, as in arranging books son a shelf. The word classification is no more satisfactory, for it denotes arranging or ordering things by class and is a term better reserved in archival science for the process of organizing active records. By contrast, the essence of archival arrangement is the identification of the natural accumulations of archival documents or records which take shape during the process of their generation

The archival classification had evolved in the past 30 years. However, if we compare with KO field, archival classification is still in an embryonic activity, because the major standardization in the areas started only in the past 10 years, making difficult to the recognition of organization patterns in arrangement/classification concepts and most important the study of classification itself. Terry Cook (1992, p.58) says that the lack of standards "when reflected in national networks of archival description could be extremely misleading to researchers."

So, the archival classification urges for standardization and, most important, the study of how these standards behave in archival classification schemes, and perhaps is exactly where the KO theory can help the Archival Science.

Its obvious, that the archival classification have major differences compared to the knowledge organization, most of them are because the type of documents that we archivists works, organic and administrative documents and records and our main principal – the provenance – this concept that bound us in a field is the one responsible for the construction of a new kind of classification schemes, what we can call as funcional-organic classification.

Archival classification schemes are more general. If we compare to the library classification schemes, the archival provenance works in two ways:
There are two aspects to the structure of archival fonds. On the one hand, archival documents are
systematized according to the way their agent of provenance organizes or structures its activities. This
external structure of provenance identifies and explains the various administrative relationships governing
the way organizations and persons conduct their business, which in turn governs the way they create and
maintain their archives. On the other hand, every archival fonds also has a documentary structure
established by the way the documents are ordered during the conduct of affairs. This internal structure of
provenance identifies the relationships among the documents as they were organized by the agent
accumulating them (Eastwood, 1992 p.4)

These pre-establish organization generates more problems than we can solve. As all
institutions have its inner and outer provenance makes, the classification data almost
impossible to be applied in different institutions and it is difficult to see the classification as
a combined idea possible to apply as standard in any institution.

The KO area can help us in some ways in this problem as, for instance, studies as the ones
carried out by Tennis (2010, 2011) in how some classes change or not in the past and in
now days in the CDD, we have the same kind of "movement" in archival classification, as
the classification scheme works as a simulacrum of its development time.

For instance, the past understanding of arrangement in the Jenkinson, and Muller, Faith
and Furin manuals is mostly related to defunct or closed funds, making the arrangement
work easier to do and to theorize about it. These kinds of funds still exists in the archival
institutions, however, we cannot see the then as a rule, because in day-by-day
documentation most of the funds still open and still receive new series and documents,
making the arrangement/classification more difficult to work. These movements happen in
the theory of the archival field but more important than that, change the face of how we do
it in practice.

The work between these two fields can help the growth of both areas, in archival science
helping with the standardization of its process, and in KO, as a counterpoint to its
classification view.

Aims and Methodology

As general objective, this paper intended to offer contributions on Archival Classification
in the light of the knowledge organization theory. As specific objectives, we perceived the
similarities and divergences between these two areas in what is related to the classifications
schemes thinking in a theoretical contribution for the Archival Science.

As methodological procedures, the study began by a bibliographic analysis on Archival
Classification, and Knowledge Organization related to the classification as a practice and
theory.

Conclusions

As was discussed earlier, the Archival Science classification theory is part of a key
activity for the organization of digital and physical records, within a public or private
institution. In the present time, the classification activity has been redesign, however, the
classification scheme or even the classification concept are still not clear, or how we can do
the classification in Archival Science and how we define this activity.

In order to discuss the difference and similarities between two fields we presented some
key aspects of the historical development from the archival classification as an activity and
how things are in the present time and also some conceptual of the knowledge organization
field.

Based on that we perceive with this article, we pointed out some theoretical possibilities
to the Archival Science in the light of the Knowledge Organization in what’s related to the
epistemological foundation of the classification schemes.

The contribution for the KO field in what is related to the archival classification
perspective is that we need to comprehend the concept of knowledge in a broad sense, that
not only the scientific knowledge can be reliable field. The administrative and organic documents can be part of this universe. The documents that need to be preserved and work as evidence of something also can be taken into account, and may work as a different view or a counterpoint not only for the knowledge concept but also to the scientific studies in knowledge organization.

References
Pekka Henttonen - University of Tampere, Finland

Diversity of Knowledge Organization in Records and Archives Management

Abstract
Different cultural, administrative and linguistic areas make knowledge organization diverse in records and archives management. The paper suggests that there are four salient differences: what knowledge organization systems (KOS) there are to achieve the goals of records and archives management, what is the moment when they are applied to organize information, what is the granularity of actions that the KOS supports and how the KOS is combined with organizational work and records processes.

Introduction
Knowledge organization is at the heart of any library, archive and museum. In this paper I examine the role of knowledge organization systems (KOS) in records and archives management (RAM). The concept of KOS is understood very broadly (about KOS, see Hjørland, 2007; Hodge, 2000): it includes all methods that are applied to make information manageable and usable either conceptually or physically: physical arrangements as well as conceptual tools like classifications, taxonomies and indexes. The purpose is to explore differences in records and archives management across cultural, administrative and linguistic areas—i.e. “RAM cultures”—in the light of some examples. However, no attempt is made to systematically define what constitutes a RAM culture. I argue that the world of knowledge organization is in records and archives management diverse compared to libraries, for instance.

Multiple KOS for Multiple Purposes?
One of the differences between libraries and archives is that in libraries knowledge organization serves primarily information retrieval. Therefore, development of search engines has led to a crisis of bibliographic classification (Hjørland, 2012). Also in RAM the role of classification has diminished. In electronic environment users prefer to search using metadata fields instead of classification scheme (Gunnlaugsdóttir, 2006; Singh, Klobas, & Anderson, 2008, p. 239). Nevertheless, in RAM classification has retained its role, because information retrieval is only one of its many purposes. In RAM literature we find several ideas about what a KOS can or should be used for.

Firstly, the organizational KOS is used to identify areas from which information is needed: the KOS serves as a basis for developing key components in a records management program: considering what documents might be needed to capture in a records management or archival system for evidential purposes. (Kennedy & Schauder, 1998, p. 115). Making functional context of records visible is another function of records and archives management KOS. The business context in which the record was created, received and used should be apparent in the record (including the business process of which the transaction is part, the date and time of the transaction and the participants in the transaction). The primary mechanism employed for this purpose is records classification. It links records to business. Grouping of records together reflects what was done. Classification enables assertions about the authenticity of records by providing context. (Ashton & Reed, 2010; DLM Forum Foundation, 2011, p. 70.) Classification serves as a basis for defining and assigning security levels and controlling sensitive information and for separating important documents from less important (Kennedy & Schauder, 1998, p. 115).
All the functions require knowledge organization in some form. One difference between RAM cultures can be found when we look at what KOS there are and whether one KOS serves more than one purpose. In Anglo-Saxon countries there is a KOS (retention schedule or disposal authority) for assigning retention periods and another (record plan) for classifying records. In Australia both are developed from and aligned with the business classification scheme. (National Archives of Australia, 2003.) In contrast, the traditional Finnish solution was to have three independent classification schemes in RAM: one in records management plan (for guiding retention, disposal, and access), the second in registry (serving information retrieval from registered records), and the third in archives (defining the structure of archival record series and serving information retrieval from the archives). Today it is recommended that the same classification scheme is used in all three. (Arkistolaitos, 2007) A third variant is found in the latest European specification for electronic records management systems (Moreq2010): besides classification and disposal schedules, it recognizes need for “aggregations”. An aggregation can be a traditional file, but it can also be a database or an online library, for instance. An aggregation may contain records that fall under different business classifications. In that case aggregation groups records together, but each record is still classified separately. Alternatively, records can get their classification from the aggregation. (DLM Forum Foundation, 2011, p. 70). Thus, in a MoReq 2010 system there are several layers of knowledge organization systems operating simultaneously.

Time Dimension of Knowledge Organization

A second major differentiating factor in RAM cultures is found when we examine time dimension: what KOS there are during record life span and what is the moment when they are applied to organize information. The time dimension is not irrelevant. There is a difference in who creates and applies a classification scheme and which is first, information or its classification. Some types of KOS are easier to create—and perhaps to use—at a particular moment of record life span. An archivist who works outside the context of record creation may find it hard to identify the original function in which the records were created. Some functions overlap (for instance, should wages records be placed under the personnel function or the finance function?) and different archivists may discern different functions in the same collection. (Williams, 2006, p. 78.) Therefore, for an archivist creating a functional classification scheme afterwards and applying it retrospectively to records may not be a viable option. On the other hand, if the functional classification scheme is created before the records and employees in the organization do the classification, the process should be easier. Even in this case it may be difficult for employees to identify the macro level functions that their work is part of (Alberts, Schellinck, Eby, & Marleau, 2010; Foscarini, 2012; Orr, 2005).

RAM cultures have both short and long term differences in the time dimension. Even at a short term in organizations (ignoring archival dimension) there are two different records management traditions. The first records management tradition is characterized by registry systems and pre-action aggregation and routing of documents. This tradition was exported from Britain to many Commonwealth countries, but it has now disappeared from many of them (Tough, 2006). On the other hand, in electronic environment registration has again become more important (Rankin, 2006). In Nordic countries registries have never lost their central position in RAM.

In registry systems, registers—books, cards or other media—are used to list the receipt and movement of the recordsof an organization while they are in current use. Registers (or journals) are an index to the records. The control over records is established at the time of their creation and receipt, before they have been processed or acted upon by the person.
responsible for handling a particular matter. Hence, most registry systems represent a highly centralized approach to active recordkeeping, in which one central registry office serves as the central recordkeeping repository for an entire organization. (Kennedy & Schauder, 1998; Maclean, 1959; Stephens, 1995.)

The second records management tradition is American. The American tradition is, in contrast, based on individual action followed by post-hoc filing (Tough, 2006). The first stage in the information life cycle is “creation” in which records are created or received. During the creation phase the records are usually managed by employees themselves. Consequently, during this phase filing systems “tend to reflect the individual paperwork styles of the individual.” Only in the next phase—when the records are moved into a more central departmental location they come under the control of records manager and departmental filing system. (Penn, Pennix, & Coulson, 1994, pp. 13–15.) Thus, the American tradition leaves more room for employees’ personal information management than the registry tradition.

Both traditions have survived in electronic environment. Besides Germany (Miller, 2003) registry tradition is common in Nordic countries. A modern Finnish public sector electronic records management system (ERMS) has always a registry component (Henttonen, 2009). In addition, electronic environment has also created its own variations of records management’s time dimension: a “front-end” ERMS captures records at or soon after creation whereas a “back-end” ERMS solution is to capture records when they “fall off” an electronic document management system (Rankin, 2006).

When we look record life span as a whole, there is another major cultural difference: how do we see the relationship between records and archives management? Are there separate phases in record life span or are they, on the contrary, inseparable parts of one “continuum”? These issues are in the heart of the two models conceptualizing record life span: North-American life cycle model and Australian records continuum model. (for an introduction to the models, see e.g. Gilliland–Swetland, 2000; Tough, 2006; Xiaomi, 2003.) In the life cycle model a record goes through phases in which it is managed by different professional groups (records managers and archivists) and used by different user groups (organization’s employees and researchers) for different purposes. In contrast, the records continuum model makes no distinction between “active” and “historical” phase of records.

Although the models do not say much about knowledge organization per se, it is easy to see that the transfer of records from one phase to another in the life cycle model (in particular, from an organization to an archival institution) offers a moment where information can be reorganized to meet the needs of new users. In contrast, in records continuum model this option does not exist: records may simultaneously have practical and historical value and they do not flow in one direction only. (Tough, 2006). Thus, in the life cycle model a KOS might be replaced by a new one. In records continuum thinking this is not viable, but there is the possibility of adding new layers of data when records pass from one domain of action to others, for example, out of the domain of the record creator into the broader domain of the workgroup and then on to the organization. (Reed, 2005.) In records continuum model a KOS is thus complemented and enriched, but not replaced. ¹

A related issue is the interpretation of the principle of provenance. Whether replacing the original KOS with a new one at archival stage is permissible depends on how we interpret the principle of provenance. The principle states that records / archives of the same provenance (fonds) must not be intermingled with any other provenance; a provenance being “the agency, institution, organization or individual that created records/archives in the

¹ In electronic environment it is also possible to have several views and multiple classifications to same information simultaneously (e.g. Bak, 2012).
conduct of its business prior to their transfer to records center/archives". (Walne, Evans, & Himly, 1984, pp. 130, 134 and 143). In Nordic countries, like in Sweden, the principle is extended to include also "registry principle" which states that also the original order (i.e. the physical KOS with its information retrieval system) should be retained in archives (Gränström, 1995, pp. 21–22). However, not everyone agrees with this.

The way the principle of provenance is interpreted has a connection to the records management tradition that is followed in the country. Registry system makes it easy to establish the "original" order and to return documents to their "right" places. (Lindh, 1994, p. 196). Thus, in countries following registry tradition the original KOS is usually preserved intact also in the archival institution. If, on the other hand, registry tradition has not been followed, there perhaps is no (at least identifiable) KOS which could be retained in archives (Petillat, 1994, p. 183). In that case introducing a new knowledge organization system at the archival stage can be a necessity.

**Granularity of Records Management Actions**

Thirdly, records management actions may take place at different levels of granularity. Also this makes RAM landscape diverse. Especially in paper environment it is usually practical to operate above the level of individual records. Practically all actions—information retrieval, contextualization, setting access restrictions, disposal, and registration—can be targeted to a group of records instead of a single record. Thus, when a KOS is designed one must consider not only what actions the KOS should support but also the scope of the actions. RAM cultures differ in this respect. Sometimes the focus is on individual records, sometimes on groups of records; containers, files or record series.

For instance, disposition—usually destruction—may take place at the level of individual documents, files, or series. In Anglo-Saxon countries the records management apparatus is geared towards purging files or records series. For instance Montaña (2010, p. 40), who writes primarily to American audience, assumes that "as a practical matter" disposition will occur on the container level—a file folder, box, tape, disk, or whatever: anything else would be "too costly and time-consuming" because it is unrealistic to expect users to open the container and select individual items for destruction. Therefore, the lowest level of classification should point to the container, and not to part of its contents. (Montaña, 2010, p. 40). Electronic environment, on the other hand, makes it possible to purge individual record types as well as folders or whole branches of classification. (Rankin, 2006.)

The Finnish practice is different. In Finnish public administration disposal is based on combination of record type (e.g. "contract", "memorandum", "letter", "ledger") and business classification scheme: the place of a record in business classification scheme together with its type determines how long the record will be retained. Therefore, in the functional classification the lowest level consists of record types which are generated in the function. Consequently, different record types in the same file may have different retention times. A similar approach is chosen also in Finnish SÄHKE specification for electronic records management systems (Arkistolaitos, 2008).

**Combining Knowledge Organization Systems with Work and Records Processes**

Fourthly, RAM cultures differ in what is the expected role of user groups in records processes during the record life span. Generally, during a record life span there can be at most four broad groups of users: two groups of specialists (records managers and archivists) and two groups of amateur users (organizational employees and archival researchers). The groups have different goals and different amount of background knowledge when they access a knowledge organization system. This makes some types of KOS more appealing to some groups than the others. The ways the user groups are
supposed to use a KOS is not the same across RAM cultures. Sometimes a KOS is intended primarily to specialist users. For instance, Australian records managers perceive classification scheme as a records management tool for grouping records for destruction. Thus, users are not supposed to have interest in it. (Singh et al., 2008.) In contrast, a Finnish textbook advocates business classification scheme as a general tool which is used by all employees even in non-records management tasks, like in introducing new employees to the organization (Lybeck, 2006, p. 80).

An important issue is the distribution of work in records processes. Ultimately it sets requirements to the KOS and defines its users. An example makes this clear. In Finland records managers plan record life cycle (including its archival stage) in advance. In principle, they define record life cycle from cradle to disposal before the records are created. As a practical tool for themselves and for the organization they create a business classification scheme in which organizational functions and record types are described. This tool (known as AMS) also contains instructions for disposal and filing and it identifies sensitive information and defines possible access restrictions. What is left to organizational users is to make some minimal selections (like selecting the right functional class and record type) and then to follow the instructions. Files have no names, because information retrieval is based on registries. The title of the registry entry is in some ways comparable to a file title. It is a free text description and does not play any role in appraisal which is based entirely on the AMS. In principle, decisions about appraisal are not re-considered after the AMS is created and in principle in electronic environment disposal can be entirely automated.

In contrast, an Australian solution is to use a functional thesaurus (faceted classification system). In this system users pick keywords from the thesaurus to create file titles. Files with different retention value are identified by the titles. Decisions about disposal take place afterwards by looking at the file titles. Not all term combinations (and corresponding disposal actions) are known in advance and some appraisal decisions are made post hoc. (National Archives of Australia, 2003.)

Hence, in the Australian practice it is important that the KOS supports naming of the files and that the users are able to name the files correctly and consistently. In Finland this requirement does exist. Finnish records management is strongly based on pre-appraisal which requires that there is an enumerative classification scheme in which all possible function – record type -combinations together with the appropriate records management guidelines are defined in advance. Introducing an Australian type functional thesaurus in Finland could not take place without revolutionizing the whole idea of how record life span is managed.

Discussion

Together, the differences make it difficult to examine knowledge organization in records and archives management abstractly without considering its total environment. What is the purpose of a KOS, what KOS and user groups there are, how they are supposed to use a KOS, and what is the role of the actions taken by the users greatly varies from one RAM culture to another. A KOS which is ideal in one context for one group and for one purpose can be less optimal when it is examined in another environment from a different perspective.

A problem of the current literature is that it rarely rises above a particular RAM culture. Although a higher level of abstraction is difficult to achieve, it is needed to advance research of knowledge organization in RAM. Existing literature takes usually for granted unspoken assumptions of the environment. For instance, Williams (2006, p. 78) argues that record types should not be used as a basis for knowledge organization, because it makes
information retrieval impossible without an index to records. This is true, but in RAM cultures following the registry tradition this is not a problem: the registry serves as the index to records.

Although some aspects of RAM diversity are well-known, it is rarely discussed how deep the differences go and how they impact knowledge organization. This creates a problem of comprehension especially for novices who just enter international discussion. A name or even a short description of a KOS does not give away all its connotations. This makes it hard both to communicate your own ideas and to understand what in the international discussion might be relevant. Nevertheless, in a globalized world—and especially in research—we should be able to clearly articulate the ideas about RAM knowledge organization across borders. The first step to this direction is to be aware the diversity of knowledge organization in records and archives management.

References


Rankin, F. 2006. Implementing EDRMS and shaping the record. In A. Tough & M. Moss (Eds.), Record Keeping in a Hybrid Environment. Managing the creation, use, preservation and disposal of unpublished information objects in context (pp. 27–46). Oxford, England: Chandos.


Leonard Will - Willpower Information, Enfield, UK

The ISO 25964 Data Model for the Structure of an Information Retrieval Thesaurus

Abstract
International standard ISO 25964-1:2011 – Thesauri for information retrieval, includes a detailed data model for thesaurus structure. It is intended to provide a rigorous presentation of the elements and relationships which will not only clarify and standardise the varying and conflicting interpretations which exist but which can also be implemented consistently in automated systems. It makes a clear distinction between concepts and the terms which are used to label them, and includes other features that may be present in a thesaurus, such as compound equivalence, arrays and node labels, concept groups, notes and version history.

Introduction
The recently-published international standard, ISO 25964-1:2011 – Thesauri for information retrieval, presents a data model for thesaurus structure which is more extensive than any published previously. The model is shown in Figure 1, using UML conventions (Object Management Group, 2007). Previous thesaurus standards are summarised by Dextre Clarke and Zeng (2011), so they will not be discussed here other than to note that they did not provide a model for the structure of thesaurus data. The new model is intended to provide a rigorous presentation of the entities and relationships which will not only clarify and standardise the varying and conflicting interpretations which exist but which can also be implemented consistently in automated systems. The SKOS (Simple Knowledge Organization System) format (W3C, 2009) is designed to present KOS data in a format that is suitable for machine inferencing and particularly for use in the Semantic Web. This is largely compatible with the ISO model, but does not yet implement all its features. Discussions are continuing on possible extensions to SKOS to cover arrays, concept groups, compound equivalence and version history.

Structure Based on Concepts, not Terms
The model is based on the understanding that thesauri show the relationships between concepts – units of thought – and distinguishes these from the terms that are used to label these concepts. These terms may be in one or more languages and one term per language is chosen as a preferred term for each concept. One or more additional terms for the same concept may be recorded in the thesaurus, as non-preferred terms. This linkage of multiple terms to the same concept is another way of expressing the traditional equivalence relationship between terms normally indicated by the tags USE/USE FOR, although the model does also show that relationship, for compatibility with existing systems. It additionally provides a “role” attribute which allows the nature of the relationship to be specified if desired, for example that the relationship between a preferred and non-preferred term may be abbreviation/full form, formal/informal, obsolete/current or scientific/popular. It was thought unnecessarily complicated to provide for such relationships between one non-preferred term and another.

Multilingual Thesauri
It is sometimes argued that some concepts which occur in one language do not occur in another, so that both language versions cannot use the same structure. This argument is invalid, because although a language may not have a specific term to represent a concept, the concept can still exist. In many cases a term is sufficient to identify a concept unambiguously, but the general principle is that concepts are not defined by the terms used to label them but rather by their place in a hierarchy and by scope notes which specify the criteria for distinguishing them from their parent concepts. For this reason the model specifies that each concept should have an “identifier” which is independent of the terms
which may be used to label it. If a language does not have a convenient label for a concept, it may be necessary to use a compound term, a "loan term" from another language, or the term which most nearly matches the concept, with a note to specify how its meaning is restricted or extended for use within the thesaurus.

**Fig. 1: The ISO 25964 data model**

Boxes in this diagram indicate "classes" in UML terminology. The diamond symbol represents an "aggregation", i.e. a "has-a" relationship. The triangle symbol indicates "generalization", i.e. an "is-a" relationship, linking a general class with more specific types of that class, which inherit all the attributes of the parent class. Normal associations (without a diamond) indicate a relationship where each instance of one class is linked to an instance of another class.

Symbols such as 1, 0.* and 1.* indicate “one”, “zero or more”, or “one or more occurrences” respectively.
Compound Equivalence

A more complex case than the simple USE/USE FOR or “preferred/non-preferred” term relationship is that of compound equivalence, where a compound concept, such as coal mining, does not exist in the thesaurus but has to be expressed as a combination of two or more simpler concepts which are there. This is shown symbolically as:

coal mining
USE+ coal
USE+ mining

with reciprocals such as “coal UF+ coal mining”. Because the complex concept is not in the thesaurus, there is no provision for recording its attributes or giving it a scope note – it has to be interpreted from the scopes of the component concepts. As a thesaurus is normally
used for post-coordinate indexing, the indexer would assign the two terms coal and mining to a document without expressing any relationship between them. A searcher would be expected to construct a search statement combining these terms with a Boolean AND operator. In the terminology of set theory, coal mining applies to the “intersection” of the set of documents that deal with coal and the set of documents that deal with mining. On the other hand a compound concept may apply to the “union” of two or more sets of documents rather than their intersection. Although ISO 25964 does not specifically deal with this case, it is generally better for the thesaurus builder to add such a compound to the thesaurus, showing its components as narrower concepts, rather than expressing it as a compound non-preferred term. For example, rather than

- fossil fuels
- USE+ coal
- USE+ natural gas
- USE+ petroleum

it is better to have

- fossil fuels
- NT coal
- NT natural gas
- NT petroleum

Hierarchical Relationships and Transitivity

Hierarchical relationships between concepts are included in the ISO25964 model, and the traditional symbols such as BT/NT are retained for consistency with current practice, although these are to be interpreted as meaning “broader concept/narrower concept” rather than “broader term/narrower term”. There is provision for each relationship to be specified by an optional “role”. This can be used to distinguish the three types of hierarchical relationship: generic, partitive and instantial, (symbolised as BTG/NTG, BTP/NTP and BTI/NTI) and even to subdivide these further if required, but in a way which allows the distinctions to be ignored by systems which do not use them.

The first level of distinction is important in automated systems and for compatibility with ontologies, where it is necessary to recognize whether a relationship is transitive or not, i.e. whether the relationship holds between concepts which are related hierarchically but where one is not the direct child of the other. A hierarchical chain in which all the relationships are generic/specific will maintain transitivity, but if it is mixed with whole/part relationships it will not.

For this reason, among others, the standard recommends that partitive relationships should normally be used only in a few specific cases: disciplines or fields of discourse, geographical locations, systems and organs of the body and hierarchical social structures. The first of these could be interpreted as generic in any case – is physics a “kind” or a “part” of science? Geographical locations is a special case because the concepts have proper names which label individual instances rather than classes, so that a generic relationship is not possible. This is different from the instantial relationship, which is used to show that an instance is a member of a class, so that we have

- countries
  - NTI India

but

- India
  - NTP Karnataka
Top concepts

Each concept can have a pointer linking it to the concept at the top of any hierarchy in which it occurs. These top concepts can be facet names, for example, and this can facilitate browsing by clearly indicating which facet a concept is in. It can also be used for validation, because hierarchical relationships are valid only if the two concepts are in the same facet. In addition, a concept can have a Boolean (true/false) attribute to indicate whether it is a “top concept”. This can be useful in producing a list of top level concepts from which to start browsing.

These links and attributes are strictly speaking redundant, because top concepts could be identified by navigating up the hierarchy until no more broader concepts can be found, but as this would use substantial processing resources it will generally be more efficient to store the information rather than determining it every time it is needed.

Associative Relationships

Similarly, associative relationships can optionally specify the nature of the relationship, such as cause/effect, process/product, or person/discipline, while allowing these all to be treated as the catch-all “related concept” (RT/RT) when necessary. This allows a thesaurus to come closer to the approach taken in ontologies, where the nature of every relationship is specified.

Arrays and Node Labels

Groups of sibling concepts, which have a common parent concept, may be organized into arrays, introduced by node labels. These are an important and helpful feature for navigation, browsing and selection of terms when hierarchical displays of thesauri are presented to users, and many existing systems do not handle these well. The order in which concepts are displayed within an array may be different from the alphabetical order of preferred terms, perhaps following some inherent sequence such as number, size or age. This is indicated by an attribute of the array entity called “ordered”, which may be true or false. Node labels, which normally contain a characteristic of division (such as “by age” in the node label “people by age”) do not represent concepts and do not have hierarchical or associative relationships with concepts. They are not preferred or non-preferred terms, although the limitations of some thesaurus software force them to be treated as such.

Arrays may be nested, so that any array may contain subordinate arrays as well as individual concepts. For example this extract from the Art and Architecture Thesaurus (2012) shows arrays introduced by the node labels <nails by form: head type> and <nails by form: point type> as subordinate arrays within the array with the label <nails by form>.

nails
  – <nails by form>
    – – cut nails
    – – helical nails
    – – hook nails
    – – <nails by form: head type>
      – – – double-headed nails
      – – – flat-head nails
      – – – headless nails
    – – <nails by form: point type>
      – – – barbed nails
      – – – blunt nails
      – – – chisel nails
Concept Groups
Many thesauri group concepts into subsets, often discipline based, called “themes” (Eionet, 2011), “microthesauri” (Unesco, 2012), “domains” (Eurovoc, 2012) or “groups”. The box in the model called “concept group” provides for such groups. The concepts within such a group may or may not have any hierarchical or associative relationship with each other, and may be drawn from distinct hierarchies or facets of the thesaurus, such as activities, people, places or things. Concept groups may be nested, and may have a scheme of notation distinct from that used for concepts or arrays, thus providing the possibility of a classified arrangement which complements the generic hierarchy of the thesaurus itself, as in a “Thesaurofacet” (Aitchison, 1970) or “Classaurus” (Devadason, 1985).

Notes and Attributes
The model provides for notes of various types to be associated with concepts and terms, as well as allowing the addition of custom notes to cater for the particular needs of special applications. Any note may contain references to other concepts, and the model allows these to be stored as links rather than text, to avoid the common problem of references not being updated when changes are made. In addition, many of the boxes in the model include several attributes, and where possible these have been drawn from other standard schemes; many of the attributes of the thesaurus as a whole, for example, are those of the Dublin Core Metadata Initiative (2012).

Version history
There is provision for attaching a version history to a thesaurus, recording the various versions which have been made available, and for each of these showing what distinguishes that version from others and whether it is still current. The attribute “this version” in each of the entries in the list of versions is set to “true” for the version in hand. Dates of creation and modification can also be attached to each concept and each term.

Part 2, Mapping, Coming Soon
The model given in ISO 25964 is for a single thesaurus. It may be multilingual, but the structure of concepts does not differ between languages. Mapping, or the creation of relationships between two or more thesauri or other types of knowledge organization schemes, will be discussed in Part 2 of the standard, currently in draft. To extend the model to cover such mapping would require models for each scheme to be shown side-by-side, with relationships between the concepts of one and the concepts of the other.

Acknowledgements
I am grateful to Stella Dextre Clarke, leader of the ISO 25964 project, for helpful comments on an early draft of this article.

Permission to reproduce extracts from ISO 25964 is granted by BSI. ISO standards can be obtained from the ISO store at http://www.iso.org/iso/store.htm and British Standards can be obtained in PDF or hard copy formats from the BSI online shop: www.bsigroup.com/Shop or by contacting BSI Customer Services for hardcopies only: Tel: +44 (0)20 8996 9001, Email: cservices@bsigroup.com. ISO standards are also normally available from national standards bodies.

The data model in diagrammatic form is publicly available on the web site for the ISO25964 project, at <http://www.iso.org/iso/25964/>. An XML schema intended for use when exchanging thesauri in whole or in part has been derived from the data model and is on the same site together with related documentation and a test document illustrating how a typical thesaurus conforming to the ISO 25964 data model can be serialized in an XML format.
References
Wieslaw Babik - University in Cracow, Poland

A Faceted Classification Scheme of Cartographic Materials: Problems of Construction and Use

Abstract
The goal of the article is to present the problems which occur in the construction of new version of Faceted Classification of Cartographic Materials, developed in the early 1990’s, with focus mainly on map content classification, as well as the methods of problem resolving. It shows application of facet analysis to one domain/document type. That is an example of special classification and of specific application of the philosophy of faceted classification to cartographic materials. The completed classification constitutes a design of a special faceted classification language, applied for indexing and retrieval of cartographic materials in a documentary information system, with potential for application in computer databases on cartographic materials and in the Internet.

Introduction
The issues of proper indexing and retrieval language application in formal and content description of cartographic materials, followed by search for information in cartographic collections, have been discussed for some years. The results of such discussions include the diversity of the currently existing classification systems for such materials, although they do not match the new information environment and constantly changing and diversifying information needs of document users. The use of a majority of material classifications reveals considerable losses of possibly important information, which reduce the degree of fulfilment of the user’s needs, and, consequently, user satisfaction. Such systems, mainly based on traditional methodological foundations, are supported by the structures of cartographic collections, and they hardly take into account their distinctive features that can constitute relevant elements for the collection users.

In this paper, I will only discuss the topic in respect to maps, however, taking into account the fact that cartographic materials also include plans, globes, atlases, aerial photographs, satellite images, cartograms, relieves or globe segments etc. Until recently, maps concerned only the Earth which is reflected in old map definitions. Dynamic space exploration caused that we have space maps as well, e.g. the NASA Solar System Atlas which shows maps and photographs of the solar system planets. Maps which are one of the products of cartographic activity are considered to be one of the oldest forms of information visualisation. It is quite a new development to look at the problem of such materials in the light of information and knowledge organization within them and of information about them. That idea will dominate my further considerations.

Multi-Dimensional Information Space of Cartographic Materials
To present that space, I will characterize the cartographic materials by distinguishing their features. Maps have their contents or topics. They also have specific forms. They are distinguished by presentation of objects, phenomena, and processes depending on topography and chronology.

Distinctive features of particular types of cartographic materials are indicated in Table1. In comparison to (Babik, 1991) it is modified.
Table 1: Distinctive Features of Cartographic Materials

<table>
<thead>
<tr>
<th>Types</th>
<th>Features</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Map, Plan, Atlas</td>
<td>Type of material</td>
<td>Subject matter</td>
</tr>
<tr>
<td></td>
<td>Form of publication</td>
<td>Surface/Territory/Place</td>
</tr>
<tr>
<td></td>
<td>Scale</td>
<td>Date of the situation presented</td>
</tr>
<tr>
<td></td>
<td>Method of presentation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Format</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Authors/Authorship</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Title</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Editor/Publisher</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Designation, Purpose</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Colour coding</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Date of issue</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Language</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cartographic projection</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Co-ordinates</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Orientation</td>
<td></td>
</tr>
<tr>
<td>Globe</td>
<td>Type</td>
<td>Subject matter</td>
</tr>
<tr>
<td></td>
<td>Dimensions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Manufacturer</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Designation, Purpose</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Date of issue</td>
<td></td>
</tr>
<tr>
<td>Aerial photographs</td>
<td>Type</td>
<td>Area</td>
</tr>
<tr>
<td>Satellite images</td>
<td>Scale</td>
<td>Surface/Territory/Place</td>
</tr>
<tr>
<td></td>
<td>Photograph technique</td>
<td>Date of the situation presented</td>
</tr>
<tr>
<td></td>
<td>Physical features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Author/Authorship</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Date of taking</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Date of developing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Colours</td>
<td></td>
</tr>
</tbody>
</table>

Particular map elements have different importance for the users. Geographic coordinates determine clearly the cartographic material, indicating the map’s location and scope. They provide basic information of identification nature of maps and other cartographic materials. Cartographic projection (equal surface or equal distance projection) allows us to establish the usability of the map for a specific purpose. The type of projection offers information about the possible map applications. The information on the scale (number, denomination, linear scale) allows evaluation of the usability of the cartographic materials. The information on the currency of data is found in the date. The dates can be diverse because of various drafting process. Orthophotomaps are the quickest to develop, and the topical maps take the longest time to prepare. Information about the medium is also important for the user.

Other essential data include: legend, emblem, authors, publisher, etc. The legend constitutes the key to learn the map contents. The emblem concerns mainly topographic maps. It clearly defines the map and constitutes the basis of search. The map title should allow for direct map retrieval. Practically, however, the titles are not clear enough. Aerial photographs and satellite images hardly have any titles. Authorship concerns mainly topical maps, while the name of the publisher is often the only indication of authorship.

Geographic Information System (GIS) generate various data, e.g. topographic data supplemented by socio-economic and geographic information, as well as access to old digitised maps. Consequently, such options are available to users and the use of traditional collections is gradually decreasing. Thus, libraries face an important challenge to supplement their analogue documents with the capabilities of presenting them on displays as digital maps, with inclusion of such information on the resources in computer catalogues, as well as allowing for multiple searches in the collections. What is an essential element here is the degree of the map’s independence. The majority of cartographic
materials are composed of maps and plans included in text documents. They are not independent in contrast to individual maps.

The complex nature of cartographic communication, structures and properties of maps as the models of reality, as well as of the cartographic information and its presentation in cartographic materials, require such organization of the information language’s semantic field that would ensure the complexity of information and meta-information rendering in the aspect of both material content and form, also allowing for multi-aspectual information indexing and retrieval (Babik, 1999). Such possibilities are created by faceted organization of the indexing and retrieval language established and pioneered by the Indian scholar and librarian, S.R. Ranganathan.

S. R. Ranganathan’s Ideas as the Foundation of Faceted Cartographic Material Classification

Facetization is the basic direction of changes observed in the present-day indexing and retrieval language theory. In 1955, the Classification Research Group recommended faceted classification as a basic tool for information organization and retrieval [Broughton, 2006]. As we can infer from the previous section of this paper, cartographic materials by their nature require multi-aspectual presentation. Present-day cartographic material users require e.g. the possibility of simple search with one search window, as well as personalized and multi-aspectual reduction of search results from both bibliographic descriptions (specific for the given library collection and others), and the selected resources. Presently multi-aspectual reduction of search results is obtained by navigation, browsing or faceted search.

Ranganathan’s analytical and synthetic information organization is conducive to indexing effectiveness and information search. One-dimensional, hierarchic organization of the semantic and retrieval field gives limited possibilities of presentation of the complexity and diversity of resources. Facetization is a good method for a more effective presentation of meta-data. However, there is a disadvantage because facets \(^1\) have to be decided \(\text{in advance}\), which makes the system hardly flexible.

One of Ranganathan’s major contributions to librarianship is his method of faceted classification known as the analytic-synthetic method. The method, though never widely used in traditional libraries, influenced classification theory and is based on the idea that traditional systems set up categories that individual items fit into. Ranganathan’s Colon

---

\(^{1}\) The Indian mathematician and librarian S.R. Ranganathan introduced the term “facet” to library and information sciences. The term “facet” was borrowed from jewellery where it means a skew cut of an angular stone edge, multiplying the colour effects as a result of light breaking and splitting. A metaphoric use of that term in the theory of classification was intended to emphasize the multi-aspectual analysis of reality and documents whose elements are subjected to classification. Each field or even issue is considered many times through the prism of the adopted categorization, and, consequently, we can retrieve from them the facets which collect the elements belonging to the same category in each respective class. S.R. Ranganathan introduced the opposition of the facet, the “category”, to the classification theory. Showing functional similarity to the facet, the category served him rather as a tool of determining fundamental categories, while facets constituted “occurrences of the category in the class.” Ranganathan’s “facet” is a collection of isolates, or the lexical units belonging to the same category, identified in accordance with the same principle of division and ordered in accordance with generic relationships. Categories are thus meant for grouping of expressions which project basic aspects of the reality being organized. In the plane of expression of indexing and retrieval languages, they are the exponents of broad meanings which identify large semantic fields in the plane of contents. The facet is an independent aspect, dimension or feature of a document (cartographic material). Presently, the term is understood and interpreted differently. The facet is “one side, one plane of an object, or an aspect in case of a conceptual object. The facet is also a homogenous class of objects whose elements contain certain features and properties making them different from the elements of other classes (or facets). The facet is also the principle of identification for grouping of properties” (Woźniak-Kasperek, 2011, 177). Facets can become the tools of discovering information resources.
Classification starts with the object and arranges categories around its aspects (facets). According to researchers, the method is more flexible and precise than other systems.

Although Colon Classification produces class numbers too long for most book spines, its basic method of facet analysis is well-suited to the explosion of knowledge on the web. Some librarians believe that classified object by faceting is similar to how objects are already organized on the Internet and that it may be a good model for providing metadata (http://hlwiki.slais.ubc.ca/index.php/Shiyali_Ramamrita_Ranganathan).

The idea of faceted classification, constituting one of the best practices of information and retrieval languages, currently supported by technology, seems to have a large potential for information retrieval (La Barre, 2006; Uddin & Janecek, 2007). Faceted search allows for navigation within the multi-dimensional information space. The user obtains the possibility of unique “discovery” of documents and information contained therein on the topic identified by the user.

**Building Faceted Classification of Cartographic Materials**

The language was constructed with the application of S. R. Ranganathan’s faceted classification theory principles. The source of the idea presented here is found in the rich literature on the general theory and methodology of building classifications developed by DRTC (Documentation Research and Training Centre, Bangalore, India) and CRG (Classification Research Group, London, United Kingdom), too. The methodology of construction of the scheme was based on earlier work (Neelameghan & Gopinath, 1965), (Vickery, 1970) and (Mahapatra, 1979).

The new faceted classification is meant to be a means of increasing the scope of presentation of information and meta-information (decreasing the degree of information loss) contained in cartographic collections, which will increase the level of use of such information and satisfy information needs of users. It was a very important task, within designing a classification system, not only to establish a topical scope, but also and primarily to design the lexical system structure which essentially influences the so-called vertical order, namely, the distribution of particular terms in classification tables, with determination of the relationships between those terms, as well as the horizontal order, or the distribution of the lexical units in the information retrieval languages (Babik, 1993).

The organization of the lexical system was done by semantic categorization (facetization) method of the collected lexical material. That organization assured multi-dimensionality of the language’s semantic field, or its faceted organization. The subdivision of lexical units into certain categories facilitates the information collection organization, and it also provides effective means of correct interpretation of classification symbol meanings, which considerably affects the optimisation of the information retrieval process.

Faceted classification was treated as a method of organizing (structuring) of the semantic field of the indexing and retrieval language. The lexical sources provided the vocabularies of library catalogues and printed catalogues of selected Polish cartographic collections, the vocabularies of the classification of special cartographic materials, the cartographic terminology used in cartographic handbooks, bibliographies and literature, as well as in surveying and earth sciences. A mixed method of the language’s lexical system building was applied.

The algorithm of faceted classification system construction was following:
Faceted Classification of Cartographic Materials – System Presentation

I will present below the problems associated with the process of building a new scheme of faceted classification for cartographic information and materials, concentrating mainly on the classification of map contents.

The Classification System of the Cartographic Materials was based on six independent categories, three of substantive nature, two containing form and type and one as separate category. The six semantic categories were accepted for the reason of ordering lexical units. They were included in the two dimensions of the semantic field assumed in the designed classification. The categories are following:

- content feature dimension (CONTENT);
  - SUBJECT (S);
  - SPACE/AREA/PLACE (A);
  - TIME (T).
- formal feature dimension (FORM);
  - FORMAL/EDITORIAL (BIBLIOGRAPHIC) ELEMENTS (E);
  - PRESENTATION (P);
plus a separate category of USE (U).

It is nothing new, but they are genuine results of a specific application of facet analysis to one domain/document type. Form categories are separated from content categories. The subject category has a very general name and its priority and use is illustrated on Fig.1.

![Fig. 1: Relationships between Categories](image)

Table 2 shows the structure of the faceted classification.
Table 2: The Structure of Faceted Classification of Cartographic Materials

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Categories</th>
<th>Subcategories</th>
<th>Facets</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONTENT</td>
<td>SUBJECT (S)</td>
<td>OBJECTS</td>
<td>Geographical objects</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FEATURES</td>
<td>Geological objects</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PHENOMENON</td>
<td>Cultural objects</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FEATURES</td>
<td>Building objects</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PROCESSES</td>
<td>Natural objects</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FEATURES</td>
<td>Physical phenomena</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HISTORICAL FACTS</td>
<td>Climatic phenomena</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Geophysical phenomena</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Social phenomena</td>
</tr>
<tr>
<td>AREA (A)</td>
<td>COMMON NAMES</td>
<td></td>
<td>Geographical and political division</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ORIGINAL NAMES</td>
<td>Territories</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Relative localization</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Names of continents, countries and places</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Names of geographical lands</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Names of historical and economic regions</td>
</tr>
<tr>
<td>TIME (T)</td>
<td></td>
<td></td>
<td>Chronology</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Seasons of the year</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Relative time</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Historical time division</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Geological time division</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Archaeological time division</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Frequency</td>
</tr>
<tr>
<td>FORM</td>
<td>PRESENTATION (P)</td>
<td>SCALE</td>
<td>Size of scale</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CARTOGRAPHIC GRID</td>
<td>Nominal value of scale</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CARTOGRAPHIC METHODE</td>
<td>Type of scale</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ORIENTATION</td>
<td>Type of cartographic grid</td>
</tr>
<tr>
<td></td>
<td>BIBLIOGRAPHIC ELEMENTS (E)</td>
<td>TYPE OF MATERIAL</td>
<td>Completeness of image</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PUBLISHER/EDITOR</td>
<td>Type of image</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLACE OF EDITION</td>
<td>Map shape</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DATE OF EDITION</td>
<td>Type of information</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DESCRIPTION LANGUAGE</td>
<td>Method of presentation</td>
</tr>
<tr>
<td></td>
<td>USE (U)</td>
<td>ASSIGNMENT</td>
<td>Taxonomy</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DISCIPLINES OF SCIENCE AND TECHNOLOGY</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PRACTICAL ACTIVITY SPHERES DIVISIONS OF ECONOMY</td>
<td></td>
</tr>
</tbody>
</table>

The designed classification scheme is fitted with systems for draft notation (mixed hierarchic notation), grammar, and instructions for cartographic document indexing and user questions. The multi-level topical classification of cartographic materials allows for the synthesis of symbols. Concentrating on the structure and description of the classification’s lexical system, I assumed that the notation system is secondary in respect of lexical organization.

The semantic power of the revised classification system should be higher than that of the currently used cartographic material classification systems. The semantic power offers the capability of system to increase the degree of fulfillment of the user’s needs and his satisfaction of the use of information/document.

---

2 Semantic power means the capability of system to increase the degree of fulfillment of the user’s needs and his satisfaction of the use of information/document.
deeper and more detailed access to cartographic information. Traditional classifications state explicitly the distinctive features of cartographic materials. Consequently, they are enumerative systems. They assign each document to one semantic class. However, the newly built classification allows us to include cartographic documents in as many classes as there are distinctive features of the objects or situations being described and reflected in their retrieval characteristics. Thus, our classification allows for multi-aspectual indexing and more selective material retrieval.

**Conclusion**

In future I will present the idea of adjusting the current faceted map classification system to new technological retrieval capabilities offered by the Internet. It is very important and interesting problem. The scheme/system and the collected terminology can be used for or adopted to the needs of a content description of specific cartographic collection, as well as the description of cartographic materials or documents placed in the Internet (in the Web), and building an ontology on that field. Cartographic material can hardly be analyzed by traditional categories such as Agent or Process, too.

At the present stage of adjusting my classification to the requirements of information and knowledge organization in the Internet (Ellis & Vasconcelos, 1999), the project has been partially completed, and other works remain to be done. That is still under construction. In my opinion the study should be continued and results can be published in the ISKO periodical “Knowledge Organization”.

**References**


Constructing Knowledge Classification Scheme in Industrial Technology via Domain Analysis: An Empirical Study

Abstract
This study combined three approaches of domain analysis such as special classifications, terminological studies, and bibliometrics to construct specialized technological knowledge classification scheme of fishery industry. It also integrated specialized terminologies of industrial technology to do information retrieval. The study requested domain experts to extract technical knowledge according to a knowledge analysis process. Then, with bibliometrics that analyzes global technological development trends of Grouper and Cobia. Follow by analysis from classification scheme to present knowledge gap of the domain. The outcome fulfilled the planned study purpose and the cross-domain team cooperation process and study results were agreed by project sponsors and domain experts. The study submitted following suggestions for further studies, in combined approaches of domain analysis. They are 1) Researchers utilized industrial technological knowledge classification scheme to assist fishery industry domain experts to clarify domain knowledge scope; 2) The combined approaches will be helpful to depict knowledge development and key technology of fishery industry in Taiwan; 3) The applied methods of combined and cross-domain approaches of domain analysis will be valuable for researchers in library and information studies.

Introduction
As the result of the professional development of information technology and constant emerging of subject domains that made the boundary of domain knowledge become blurry which impact the subject classification of related knowledge and literature. To develop knowledge organization system that is flexible and easy to integrate becomes an important issue. One of the workable ways is to construct the structure of knowledge classification scheme through domain analysis.

In recent years, scholars used various approaches of domain analysis to understand particular knowledge domain that had been increased volume of related researches. For example, librarians tried to assist users to search for professional information resources and filtering large amount of irrelevant internet resources, they constructed subject gateways, such as AgNIC, AgriFor, AGRIGATE, and INFOMINE (Pu, Chung & Kuo, 2005). The researcher used literature analysis, questionnaire survey, and interview with domain experts to build a draft of “Organic Agriculture Domain Knowledge Classification Scheme” in 2006. In the next, as followed by a further survey and interview toward many domain experts to identify the scope and content of the classification scheme. The outcome resulted in the recognition of the most domain experts over the knowledge classification scheme; however, it still needs to improve the methodology (Chen & Yuan, 2006). As a researcher in Brookhaven National Laboratory in USA, Tanaka utilized bibliometrics to identify the authors, citations, and their academic relationships of co-publication in the Scientific Computing Group at the institution toward a period of fifty years, from 1958 to 2007; and used historical methods and terminological studies to investigate the development and the changes in research theme of Scientific Computing Group (Tanaka, 2010). Among eleven approaches of domain analysis induced by Hjørland (2002) stated that each and every one of domain analysis works for each other and without conflicts, Tennis (2003), Mai (2005),
Lee, Kim and Kim (2010) as well, provided outcome of their studies on domain analysis which in consensus suggest that it will be possible to develop more researches and applications in domain analysis via cooperation of expertise of library and information science professionals. In a more concrete fact of domain analysis application from viewpoints of Prieto-Diaz (1990) and Yuan (2003) respectively, is that the domain concepts and knowledge structure are accumulations of the integration of subject terms from the extraction of keywords, as well as from subject heading and contents of literature.

During years of 2006-2008, under the sponsorship of the Council of Agriculture, Executive Yuan, the researcher coordinated with National Taiwan Ocean University jointly conducted a research project of “Packing and Value-added in Agricultural Scientific Technology” of Grouper (Epinephelus) and Cobia (Rachycentron canadum) which combined three approaches of domain analysis, through multiple information resources, systemically, comprehensively and objectively examined and analyzed the research capacity of industrial technologies of two items, and be used to find out the key technologies and knowledge gap. A cross-domain working process of this research project shown on Figure 1, the research was in charge of classification scheme and keywords coordination, multiple information resources collection, rules of analysis, abstract charts, and information integration that involved more than 20 professionals and researchers that successfully constructed the technological knowledge classification scheme of fishery industry. After finishing the classification scheme, the researchers utilized the classification scheme to be the representation of knowledge analysis in fishery industry which was also useful to understand the development of fishery industrial technical knowledge. At the final stage, a professional discussion group was formed up that confirmed the correction and importance of analyzed data; finally, the researchers provided strategies and suggestions of reporting.

**Fig. 1: A cross-domain working process of this research project**
The purposes of this study include:

1. To form up a cross-domain working team and construct industrial technological classification scheme: Follow by information collection of literature and classification scheme resources. Utilizing domain analysis of specialized classifications and terminological studies to assist fishery domain researchers to construct and apply technological knowledge classification scheme for fishery industry.

2. Applying industrial technological classification scheme on technological knowledge analysis of fishery industry: Following information collection of literature resources, utilizing patent analysis and bibliometrics to explore fishery industrial technical knowledge development, key technologies, and knowledge gap.

3. Collecting feedbacks: Collecting feedbacks from domain experts toward technological knowledge classification scheme construction and application on knowledge analysis of fishery industry.

Methodology

This study utilized three approaches which included specialized classifications, terminological studies, and bibliometrics, among eleven approaches of domain analysis induced by Hjørland, with in-depth interview plus domain experts’ discussion; and conducting analysis according to following industrial technical knowledge construction process:

Stage I: Identify topics for empirical study and scope of industrial technical knowledge, i.e. to confirm the definition and scope of the domain regarding topics of empirical study. Follow by collecting domain expert list to form up a cross-domain working team to construct specialized industrial technological knowledge classification scheme.

Stage II: Formation and examination of the specialized industrial technological knowledge classification scheme, i.e. collecting books, literature, classification schemes, and internet resources of the fishery industry, then, constructed the industrial technological knowledge classification scheme of fishery industry and organized terminologies; Through several meetings for discussion and revision of the classification scheme drafts and terminologies; In the end of Stage II, Reviews of the classification scheme and the terms by senior experts.

Stage III: Applying technological knowledge analysis toward industrial technological knowledge classification scheme of fishery industry: According to the outcome of Stage II, collected technological keywords from domain experts, then conducted patent and literature retrieval, also obtained full-text of the retrieved literature. Experts are requested to do knowledge analysis from the full-text according to the two abstract charts to extract technical knowledge; Moreover, through senior domain experts’ examination of the content to assure the quality. Then the data and contents were put on patent analysis and bibliometrics to clarify the development and key technologies of fishery industry.

Stage IV: Representation of the fishery industrial technical knowledge gap via industrial technological knowledge classification scheme.

Stage V: Conducting experts’ discussion to collect feedbacks toward industrial technological knowledge classification scheme construction and application on knowledge analysis from domain experts of fishery industry.

Limitation of this study is the industrial technical knowledge classification scheme emphasis on practical use; it is highly related to precision and purposes of constructing the classification scheme. This study was under an approval research project that with clear goals, coverage, team members structure (included domain industry, authority, academic,
research and library and information studies), along with highly cooperation, sufficient fund support from government; thus, the classification scheme constructed by the team considered highly practical and applicable for fishery industry in Taiwan at the current stage. The complete technological knowledge classification scheme has five level classes; this study has described only level 1 and 2.

**Results**

Confimed the scope of fishery industrial technical knowledge

Two kinds of fish (Grouper and Cobia) were targeted for this study regarding the global production and marketing of R&D outcomes in current stage. The researchers divided fishery industrial technical knowledge into seven items from perspectives of production supply chain, which items were invited at least two domain experts to do analysis. Followed by providing checklist of bibliographical and database resources for further information retrieval and selection.

**Constructed technological knowledge classification scheme of fishery industry and collected terminologies of Industrial technology**

Shown on Table 1 of technological knowledge classification scheme of fishery industry, that contained seven items of fishery industry. Domain experts were required to provide key industrial technological knowledge categories and notation of each item; Library and information experts were also assist to confirm hierarchical relations of each item. A “technological knowledge classification scheme of fishery industry” was drafted. After finishing the draft, the researchers conducted expert discussion meetings to adjust and revise the structure of the classification scheme draft. Furthermore, the researchers made a “core keywords charts” which contained Chinese words, English words, synonym, related terms, and unrelated terms. And then, domain experts were required to provide terminologies of Industrial technology according to the classification scheme by “core keywords charts”, and checked iteratively many times.

**Table 1: Technological knowledge classification scheme of fishery industry**

<table>
<thead>
<tr>
<th>Class</th>
<th>Sub-Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Industry Analysis</td>
<td>A01 Productivity Analysis</td>
</tr>
<tr>
<td></td>
<td>A02 Market Analysis</td>
</tr>
<tr>
<td></td>
<td>A03 Strategy Analysis</td>
</tr>
<tr>
<td></td>
<td>A99 Other</td>
</tr>
<tr>
<td>B Breeding Techniques</td>
<td>B01 Facilities of Breeding</td>
</tr>
<tr>
<td></td>
<td>B02 Larval Rearing System</td>
</tr>
<tr>
<td></td>
<td>B03 Broodstock Cultivation</td>
</tr>
<tr>
<td></td>
<td>B04 Management of Fertilized Eggs</td>
</tr>
<tr>
<td></td>
<td>B05 Seed Production</td>
</tr>
<tr>
<td></td>
<td>B99 Other</td>
</tr>
<tr>
<td>C Breeding</td>
<td>C01 Genetic Breeding</td>
</tr>
<tr>
<td></td>
<td>C02 Chromosomal Manipulation</td>
</tr>
<tr>
<td></td>
<td>C03 Molecular Breeding</td>
</tr>
<tr>
<td></td>
<td>C99 Other</td>
</tr>
<tr>
<td>D Grow</td>
<td>D01 Land-based Grow Out</td>
</tr>
<tr>
<td></td>
<td>D02 Sea-based Grow Out</td>
</tr>
<tr>
<td></td>
<td>D99 Other</td>
</tr>
<tr>
<td>E Living Food/Feed</td>
<td>E01 Live Food</td>
</tr>
<tr>
<td></td>
<td>E02 Artificial Diets</td>
</tr>
<tr>
<td></td>
<td>E99 Other</td>
</tr>
<tr>
<td>F Diseases</td>
<td>F01 Non-infectious</td>
</tr>
<tr>
<td></td>
<td>F02 Infectious</td>
</tr>
<tr>
<td></td>
<td>F03 Control Strategy</td>
</tr>
<tr>
<td></td>
<td>F99 Other</td>
</tr>
<tr>
<td>G Products</td>
<td>G01 Product Quality and Value Added</td>
</tr>
<tr>
<td></td>
<td>G02 Supply Chain</td>
</tr>
<tr>
<td></td>
<td>G03 Product Marketing</td>
</tr>
<tr>
<td></td>
<td>G99 Other</td>
</tr>
</tbody>
</table>

Take class C (Breeding) as example to exhibit notation and keywords of classification shown on Table 2.
Table 2: Exhibition notation and keywords of classification scheme

<table>
<thead>
<tr>
<th>Class</th>
<th>Notation</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>C Breeding</td>
<td>By mating the selected varieties to improve the breeding variety.</td>
<td>—</td>
</tr>
<tr>
<td>C01 Genetic Breeding</td>
<td>By Hybridization, Back Cross or Inbreeding conducting variety selection on those targeted characters including: Growth, Disease Resistance, Cold Tolerance and Salinity Tolerance.</td>
<td>Heritability, Genetic Variation, Mutation, Fitness, Homozygosity, Genotype, Phenotype, Selective Breeding, …</td>
</tr>
<tr>
<td>C02 Chromosomal Manipulation</td>
<td>Artificially in control Chromosomal Division by Water Pressure, Water Temperature or Colchicines.</td>
<td>Triploidy, Tetraploidy, Gynogenesis, Androgenesis…</td>
</tr>
<tr>
<td>C03 Molecular Breeding</td>
<td>By mediated nucleic acid (DNA or RNA) to conduct the Transgenesis (Gene Transfer); the molecular or gene transfer is reached to breeding target including particular genes been cloned or deleted.</td>
<td>DNA Marker, DNA Sequence Polymorphism, Nuclear DNA Sequence, Micro-satellite DNA Sequence…</td>
</tr>
</tbody>
</table>

Experts Confirmation on Technological knowledge classification scheme of fishery industry

In addition to discussion meetings with 20 experts to revise the technological knowledge classification scheme of fishery industry, there were experts of the third party to do the review over 1) Completeness of coverage and properness of the Draft; 2) Properness and correctness of class name; 3) Properness of levels and sequences; 4) Properness and correctness of notations. The outcome is as follows:

1. For the whole structure: a) Balance of levels, with more equal numbers of item numbers of each level; b) Add new technology and new products for more completeness of the structure; c) Consistency of terminology.
2. Adding class: under E0205 (Functional Diets), such as Color Enhancing Diets, Induced Maturation Diets; Under F0305 (Vaccine) add Adjuvant.
3. Class level changes: uplift F0205 (Control Strategy) into F03 (Control Strategy); uplift F02050204 (Medicine) into F0303 (Medicine).
4. Revise name of class: B02 (Breeding System) change to B02 (Larval Rearing System), thus, notation become more suitable.
5. Class notation revision: B0502 (Nursery) is artificial technology, such as prevention diseases, nutrition enrichment, water quality management, feeding management, and selection techniques into two inch size fingerling. And change Fingerling into Fry, because fingerling is smaller than Fry.

Fishery industrial technical knowledge analysis via bibliometrics

After collecting keywords and classification scheme, the researchers utilized four frequently used local scientific and technology databases, “Patent Information Databases”, “Government Research Bulletin”, “Index to Taiwan Periodical Literature System”, and “National Digital Library of Theses and Dissertations in Taiwan plus a well-known internationally “Web of Knowledge”, which used to be as the literature resources. Conducting search using keywords and obtained 5,000 related records, filtered through experts group of NTOU and output with 649 advanced records for further knowledge analysis by 20 selected experts who follow the patent abstract chart and literature analysis chart. Charts were designed by researchers. The returned abstract charts and data were examined and confirmed before documenting into database of technical knowledge.


Patent application of Grouper started in 1989 by Kibun Foods Inc. toward USA (Patent Number: US4816279). There were 64 patent assignees around the world, Sun-Yat-Sen University in Guangzhou obtained highest number of Grouper patents focused on breeding.
techniques, breeding, and living food/feed of various technical knowledge categories; the
next was Chinese Academy of Sciences which focused on breeding and disease prevention
technical knowledge categories; and National Science Council in Taiwan’s patents are all
related in disease prevention technical knowledge categories. From viewpoints of
knowledge classification scheme, “Disease Prevention” was Grouper patents occupies
highest at 45%, followed by breeding and living food/feed. It indicated that disease
prevention was a focused area of Grouper patents in global fishery industry; and products,
breeding techniques, and industrial analysis occupied less. In 2003, the first Cobia patent
(Patent Number: CN1442077A) under living food/feed technical knowledge category were
obtained by Ocean University of China. In comparison with patent numbers of Grouper,
Cobia was less. The reason from domain experts’ interview, they presumed that Cobia
patent is focused on technology of net cage facility. They suggested it should be a good
reference when doing patent search afterward. From viewpoints of knowledge classification
scheme, “living food/feed” occupied highest as 56%, followed by disease prevention and
grows out. There was no patent on technical knowledge categories, such as products,
breeding, breeding techniques, and industrial analysis.

(2) Quantitative Analysis of Research Capacity on Groper and Cobia in Taiwan

The contributions of Grouper science literature from Taiwan research institutes include
National Taiwan Ocean University, National Taiwan University, and Fisheries Research
Institute. The highest is National Taiwan University which occupied 45%. Analysis of
technical knowledge categories derived from knowledge classification scheme, academic
research capacity of Grouper in Taiwan were focused on disease prevention which
occupied around 57%, others were living food/feed and breeding techniques. The
contributions of Cobia science literature from Taiwan research institutes included National
Taiwan Ocean University, Fisheries Research Institute, National Sun-Yat-Sen University in
Taiwan, and National Taiwan University. Among these institutions, National Taiwan
Ocean University contributed the highest amount of literature. Analysis of technical
knowledge categories derived from knowledge classification scheme, academic research
capacity of Cobia in Taiwan were focused on disease prevention which occupied around
29%, others were grows out and living food/feed. From above information, one can know
that the academic research capacity of Grouper and Cobia in Taiwan both gathered in
National Taiwan Ocean University; from distributions of technical knowledge categories,
disease prevention and living food/feed and related researches are advantages of Taiwan’s
research institutions.

Representation of knowledge gap via industrial technological knowledge classification
scheme

Shown in Figures 2&3, combined patents and scientific papers of Grouper, there were
247 technical documents in Taiwan, in comparison with 118 technical documents outside
Taiwan. Major distributions of technical knowledge categories of technical documents were
disease prevention, living food/feed, and breeding techniques. Disease prevention of
technical documents was more than halves among the three, disease prevention of technical
documents were more advanced because it had more technical documents in the field.
There was more literature in breeding, disease prevention, and breeding techniques outside
Taiwan, especially in breeding, technical documents outside Taiwan were more than in
Taiwan.
The researchers combined patents and scientific papers of Cobia, there were 187 technical documents in Taiwan, in comparison with 32 technical documents outside Taiwan. Major distributions of technical knowledge categories of technical documents were disease prevention, grow out, and living food/feed. Disease prevention of technical documents was highest among the three. Technology of “Grow Out” was more outside Taiwan. Disease prevention, grow out, living food/feed, and products were more than outside Taiwan while breeding techniques was less in Taiwan. Therefore one can predict that technological development of Cobia in Taiwan was more active than outside Taiwan.

Opinions from experts on technological knowledge classification scheme of fishery industry and application on knowledge analysis

After the completion of technological knowledge classification scheme of fishery industry and knowledge analysis of Grouper and Cobia, the researchers realized and summarized the knowledge gap via technological knowledge classification scheme. A discussion meeting was held to collect opinions of experts regarding the construction of classification scheme and representation of knowledge gap, which the result was conformed to the goal of research project. Following were the opinion collected from the meeting:

1. The classification scheme was not only helpful to present the focus of technical knowledge, but also helpful to see the knowledge gap of fishery industry and research in Taiwan.

2. Domain experts also agreed that the process of constructing classification scheme.

3. Through the assistance of domain experts, one can find the connections of technologies and to extract the key advantageous technology of Taiwan.

4. During the research process, researchers accumulated patents and scientific papers, through analysis and extract data from the abstract charts that formed up very valuable resources for related researchers. Combining the classification scheme and the literature resources that will be helpful for further value-added and application.

Conclusions

The researchers used domain analysis toward specialized industrial knowledge to conduct exploration, examination, and organization toward 7 industrial technical knowledge categories, 19 sub-classes, and about 200 fishery industrial technological terminologies. These terms were used to do information retrieval, organization of bibliographic data and full texts, and requested domain experts to extract technical knowledge and made two abstract charts according to an analysis process; in the following stage, the researchers
utilized bibliometrics to analysis global technological development trends of Grouper and Cobia. The result of knowledge analysis was represented via technological knowledge classification scheme of fishery industry systematically. The outcome of this study such as the cross-domain team cooperation process, three approaches of domain analysis, specialized industrial knowledge classification scheme, and application of knowledge analysis were agreed by project sponsors, domain experts and industrial businessmen in fishery.

In future, we suggest that 1) More empirical studies on various industrial technical knowledge according to the process of this study; 2) Expanding various value-added products, such as technology portfolio strategy, R&D strategic planning and publishing on the base of industrial technical knowledge category and application achievements, that will be a helpful reference for industrial businessmen; 3) The three approaches of domain analysis are focus on explicit knowledge, thus, to extract tacit knowledge through interview with domain experts and found out that partial of know-how are not contained in the existing technological literature. The process of tacit knowledge extraction is helpful to upgrade the completion of the outcome on industrial technical knowledge analysis.

Acknowledgement
The work reported here is in part supported by the Council of Agriculture, Executive Yuan under 96農科-13.1.3-科-a1.

References
**Sanskrit-English Bilingual Thesaurus for Yogic Sciences: A Case Study of Problems and Issues with Terms of Non-Latin Origin**

**Abstract**
A machine-readable Sanskrit-English thesaurus for yogic and allied sciences is being developed using Greenstone Digital Library software (GSDL) at the S-VYASA deemed-to-be university in Bangalore. This paper deals with the problems and issues that arose during construction of the bilingual thesaurus.

**Introduction: Need for this Work**
Swami Vivekananda Yoga Anusandhana Samsthana (S-VYASA) is an academic and research centre for yoga and allied sciences and has links with a number of institutions in India and abroad. Digital library of S-VYASA consists of an online catalogue of information resources – Books, theses, dissertations, reports, conference proceedings, and research papers. Full texts of research papers, theses and dissertations are accessible on their websites. Yoga is an ancient discipline and many of the early works are in Sanskrit e.g. Patanjali’s Yoga Sutras (PYS), Gheranda Samhita, Yoga Vasistha, Hatha Yoga pradipika, Shiva samhita, etc. Even today researchers publish in Sanskrit language on the subject. In recent times, however, there are books, reports and papers on yoga, its application and effects on human health in English and in some other languages of Latin origin. In this context the need for a vocabulary management tools was felt; hence the project for the design and development of Sanskrit – English bilingual thesaurus.

**Review of Literature**
A literature survey revealed that there are a number of papers on the design and development of digital information systems and knowledge organization tools (KOTs). To cite a few: Hudon (2001), Jorna and Davies (2001), Kwasnik and Rubin (2003), Landry (2004), Millstead (2001), and Molholt (2001) dealt with problems of multilingual thesaurus design and development; however most of them deal with terms of Latin origin. Neelameghan (2001), Neelameghan and Raghavan (2005), Raghavan and Neelameghan (2008) describe their work of preparing bilingual and multilingual thesauri, F-thes (Sanskrit, Farsi, English) and G-tamthes (Tamil and English). However in the humanities and yogic sciences in particular, there are only a few papers relating to terms of non-Latin origin (Neelameghan and Raghavan, 2005; Neelameghan, 2007).

**Features of the Bilingual Thesaurus**
The following requirements were kept in mind:

- It should be possible to search in Sanskrit and English;
- The software should be Unicode compliant and capable of handling Sanskrit (Devanagari) script;

---

Enable hypertext linking within the thesaurus and with external databases and tools –
dictionaries, lexicons, relevant descriptive texts online.

It should provide for equivalent terms in both languages, if not near equivalent terms or
transliterated terms.

Problems and Issues
For many Sanskrit / concept / terms exact English equivalent terms are not available;
similarly, for the recent works about yoga in English, Sanskrit equivalent terms for some of
the specialized English terms are not available. In such cases hypertext linking to the
following alternatives can be used:

- Transliterated term(s)
- Code e.g. Classification code
- Scope note or descriptive definition

<table>
<thead>
<tr>
<th>Issues</th>
<th>Sanskrit term</th>
<th>Equivalent English Term</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yoga terms in original texts</td>
<td>available</td>
<td>not available</td>
<td>Use near equivalent terms; or Sanskrit transliterated in English</td>
</tr>
<tr>
<td>Yoga terms in modern research</td>
<td>Equivalent Sanskrit terms not available</td>
<td>available</td>
<td>Use near equivalent terms; or English transliterated into Devanagari</td>
</tr>
</tbody>
</table>

User can search in Sanskrit or English; if a code is used one can give the code and can
search all the languages in the thesaurus and external online lexicons and databases; codes
can be a class numbers from a classification scheme.

Designing the Bilingual Thesaurus (Y-Thes)
The designing of the thesaurus involved the following tasks:

1. Study of the subject by examining the literature available on the subject, moving
   progressively from reference sources – dictionaries, glossaries, encyclopedias, etc. - to
   the more detailed material, such as, papers in periodicals, proceedings of seminars,
   conferences, symposia, etc.;
2. Creating a corpus of terms;
3. Identifying user needs from yoga library users profiles;
4. Selection of terms from the corpus in consultation with subject specialists and the users
   of information system;
5. Record each term in the format chosen as most convenient; The layout of the ultimate
   product is decided and the thesaurus is constructed accordingly.

Sources of Concepts / Terms
Corpus of terms - dictionaries, glossaries, lexicons - already exists. Relevant terms from
the OM database, and the F-thes glossary were transferred. A collection of over 5000 books,
peer-reviewed journals such as International Journal of Yoga (IJY) and International
Journal of Yoga Therapy, theses and dissertations, research reports were used for
identifying core concepts. Some select secondary sources were also used (see Annex 1).

Criteria for Selection of Software
The Greenstone Digital Library software (GSDL) for building the bilingual thesauri
(Kumar and Nikam, 2011) was used as it was Unicode compliant and allowed hyperlinking
within the thesaurus as well as to external sources.
### Database

The fields and tags of the database

<table>
<thead>
<tr>
<th>Tag</th>
<th>Field Name</th>
<th>Example of data entry</th>
<th>Example of data entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Title</td>
<td>अष्टाङ्गयोग</td>
<td>अहिःसा</td>
</tr>
<tr>
<td>2</td>
<td>EQS (Equivalent Sanskrit)</td>
<td>अष्टाङ्गयोग</td>
<td>अहिःसा</td>
</tr>
<tr>
<td>3</td>
<td>EQS.USE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>EQS.UF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>EQS.BT</td>
<td>योग</td>
<td>यम</td>
</tr>
<tr>
<td>6</td>
<td>EQS.NT</td>
<td>विद्वेषन साधना</td>
<td>अन्तरर्ग साधना</td>
</tr>
<tr>
<td>7</td>
<td>EQS.NT</td>
<td>यम</td>
<td>यम</td>
</tr>
<tr>
<td>8</td>
<td>EOE (Equivalent English)</td>
<td>aṣṭāṅgayoga</td>
<td>aḥiṁsā</td>
</tr>
<tr>
<td>9</td>
<td>Term</td>
<td>aṣṭāṅgayoga</td>
<td>aḥiṁsā</td>
</tr>
<tr>
<td>10</td>
<td>ScopeNote</td>
<td>The Yogadarsana of Patanjali (200 B.C.) is named Astangayoga since it consists of eight limbs (asta = eight; anga = limb or steps). In Concise Encyclopaedia of Hinduism by Swami Harsharanda.</td>
<td>Non violence Astangayoga since it consists of eight limbs (asta = eight; anga = limb or steps). In Concise Encyclopaedia of Hinduism by Swami Harsharanda.</td>
</tr>
<tr>
<td>11</td>
<td>USE</td>
<td>Ashtanga Yoga</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>UF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>BT</td>
<td>BT: yoga</td>
<td>यम</td>
</tr>
<tr>
<td>14</td>
<td>NT</td>
<td>NT: bahiraṅga sādhanā</td>
<td>NT: antaraṅga sādhanā</td>
</tr>
<tr>
<td>15</td>
<td>RT</td>
<td>RT: Yama</td>
<td></td>
</tr>
</tbody>
</table>

### Example of Display

Sanskrit term can be accessed by pressing the ‘Sanskrit term’ button in the navigation bar. This displays a list of terms in alphabetical sequence as shown in Fig. 1. Clicking on any of the displayed hyper link terms, say ‘यम’ will retrieve matching records in the pre-selected databases (See Fig. 2). English Descriptor(s) can be accessed by pressing the ‘English Descriptor’ button in the navigation bar. This displays a list of terms in alphabetical sequence as shown in Fig. 3.

**Fig. 1: Sanskrit terms (partial list) display by clicking Sanskrit term button**
Fig. 2: Display of Thesaurus in Sanskrit

Fig. 3: English descriptor (partial list) display by clicking English descriptor button

Example

<table>
<thead>
<tr>
<th>Sanskrit Term</th>
<th>English Descriptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yoga</td>
<td>Yoga</td>
</tr>
<tr>
<td>(reduces)</td>
<td>NT Karma Yoga</td>
</tr>
<tr>
<td>RT Stress</td>
<td>NT Jnana Yoga</td>
</tr>
<tr>
<td>Stress</td>
<td>NT Bhakti Yoga</td>
</tr>
<tr>
<td>(reduced by)</td>
<td>NT Raja Yoga</td>
</tr>
<tr>
<td>RT Yoga</td>
<td>NT Ashtanga Yog</td>
</tr>
</tbody>
</table>

Utility and Value

In the Yoga Thesaurus search facility, users can type a Sanskrit term using on-screen keyboard. Search using on-screen Sanskrit keyboard, for keying in the search terms in languages other than English a Unicode compliant keyboard is necessary. The term(s) may be dragged from the list and dropped in the search box if necessary. The thesaurus is useful in domain mapping and for filling up the gaps in the vocabulary. In addition the possibility of using the thesaurus to develop a depth classification scheme for the domain is being explored.
Acknowledgment
We thank Ms. S.K. Lalitha of SRELS for help in the installation of GSDL.

References


Annex 1
Abstract
The leather shoes and artifact production chain is very competitive in Brazil. It is important to develop strategies to identify, locate, organize and retrieve information to make this chain internationally competitive. In this context, we developed a specialized system with information used in competitive intelligence - InfoSIC. This paper discusses the controlled vocabulary developed for InfoSIC.

Introduction
The institution representing the Brazilian production supply chain of leather, footwear and artifacts requires information that is suitably grouped and classified as technological, commercial, managerial and marketing information. InfoSIC (Competitive Intelligence System), an essential element of the footwear manufacturing environment, can be understood as an extension of resources that offers a broader and more strategic outlook to the actors in the supply chain in their country and perhaps in the world (Battaglia, 1999). The products and services that SIC provides to its target audience – actors in the chain – enables a more comprehensive understanding concerning the situation of their business, market trends and evolution process as a business group. It was perceived that this knowledge, in addition to being economically important to the group, is also socially important to the country as it is one of the key factors for rational decision making, and which also provides information to efficiently implement changes that can guarantee the survival or maintenance of a business and also enable a sustainable competitive advantage, quite necessary in current times.

With this brief preface, it can be stated that the purpose of InfoSIC is to centralize data and information related to this wide-ranging area, such as: data registration of companies in the industry (all those in the production chain); news, books, theses and dissertations, patents, technical standards; foreign trade data and statistics, domestic production numbers, currency quotations; scientific articles, various reports, surveys, and additional information of interest to the links in the chain. These elements are grouped into a single technological environment in order to store this information in an organized manner as well as to identify the inputs - through this organization - in the knowledge production of the production enabling to prepare information / intelligence products for decision chain, thus making by any of its users. It was with this primary function – information management – that the InfoSIC system was developed (Information System for Competitive Intelligence).

As InfoSIC is formed by a group of actors that generate, search and share data and information related to the economic sector in question, it prioritizes any one link of interest that is part of the productive chain in Brazil. Thus, InfoSIC includes: organizations representing the productive chain links of leather, footwear, components and artifacts from suppliers to end customers, service providers, government agencies, research institutes,
universities, industries of the chain, unions, experts, businessmen and others who participate directly or indirectly in the demarcation movement of the area.

The InfoSIC system can be defined as a technological tool that services national entities, unions, associations or other organizations that represent the production chain (individuals and legal entities), the actors that feed the system with information and knowledge, reciprocally providing such insertions and empowering the production of an updated overview of the segment as a whole within this collective dynamics.

This system is virtually located in a Portal - called SIC Portal (web site), where the actions are directed to disseminate information and intelligence reports, providing its users access to the updated data which is produced in their interaction. (Dias, 2001).

Some of the entities that support the development of the sector in Brazil are: SEBRAE, Apex-Brazil and ABDI.

InfoSIC is characterized as a dynamic network that feeds on the web of information produced by the elements of this component and provides, through its democratic intervention nature, the suggestions from each member of the link, while allowing the right to use its collection. Its collection mechanism is facilitated by mutual sharing, and is also more accurate in how it processes and stores data, since it meets the demands required by the activities, mapping its acquisitions or searches.

Within InfoSIC, the actors are classified as: intelligence centers (NIC) that represent a productive pole or a representative entity of the links, forming what is called Productive Chain of Leather, Footwear and Artifacts, and using this to individually identify the nomenclatures NIC POLO and NIC ENTITY that are serviced.

In addition to these NICs, the system receives technological support from UTIC (Technology Unit for Competitive Intelligence), located in the productive pole of Jau/SP, Brazil, and also from the governance of a managing committee consisting of two representatives of each link in the productive chain.

The data entry modules of InfoSIC are: Registries: Corporations, Companies in the Chain, Other Organizations, Individuals, Work Posts, Products of the Chain, Brands, Digital Library, Articles, Brochures, Books, Reports, Academic Works, Events, Projects of the Entities, Standards, Patents, Columns, Multimedia Center, News, Collections, Crafts, Shoes, Components, Leather, Machinery, Retail, Statistics, Foreign Trade, Jobs, IBGE, Quotes, Classifications, Terms, Intelligence, Unverified Information Analysis, Administration, Portal, Profiles, Users, Maintenance Requests.

In the InfoSIC system, the Classification of topics has a centralizing purpose. The instrument is responsible for devising a type of standard among the workers, documents, language used. The biggest problem in such scenario is the fact that the meaning of the product information has been constantly changed and created by the fashion industry and by the consumers.

In this paper, the objective is to discuss the vocabulary control methodology (Campos, Gomes; Dodebei, 2002) of InfoSIC. In the system the specific goal is to research and offer a specialized language to index the information within the system. The leader, the footwear and artifact productive chain is very competitive in Brazil. So, it is important to develop strategies to identify, localize, and organize or retrieve information to make this chain more competitive internationally. Thus, this paper describes the methodology, under development, for the construction and maintenance of the terminology list in the leather and footwear area, which is produced by UFSCar Team, and then implement it as a support
tool for the intelligence activities of the Competitive Intelligence System of the Production Chain for Leather, Footwear and Artifacts – InfoSIC. The methodology was carried out by cross referencing different approaches to construct languages so that they could include a mechanism for the production of dynamic information in that productive chain. The terminology list generated by applying this methodology will serve in the future as a tool for the classification, indexing and retrieval of content within the system itself. To construct this new methodology, especially directed to the sector, the following approaches/guidelines were used: international standards ANSI/NISO/2005 for the construction of controlled vocabularies; CRG – Classification Research Group, for the classification of terms; E-terms, data mining environment, based on natural language. Within the methodological approach, in order to manage the inclusion of terms, a specific field for the insertion of candidate terms was proposed, which can be used by the information classifier as well as by the searcher. With this procedure, this term is later adjusted within the internal structure of a terminology list produced to control the use that is deemed representative of the concept targeted.

With this entry of candidate terms the idea was to give the terminology list the necessary dynamism to represent information in the area of leather, footwear and artifacts. This perspective emerged after it was perceived that such terminology instrument should not merely represent the area in a static manner, but should be constantly updated and organized given the dynamic nature of the domain. As for the processes, all of the items included in the system are classified at three levels of classification, resulting in a cross-referring that will enable each concept used to describe content so it can fit into different themes. This content can be classified into the following categories: Links in the chain, Factory floor, Terminology list. As for the levels of satisfaction in terms of information retrieval, these will also be evaluated by applying a specific methodology to the different profiles that will make use of SIC. This proposal was presented to the representatives of the sector, and their demands were represented and incorporated into the methodology.

**InfoSIC Content Classification and Indexing**

The classification process in the InfoSIC environment organizes the information in the portal to establish the relationship between the records of different fields, but with the same content or additional content. Therefore classification is a very important process in the Intelligence sector, as through this representative projection, the analysts can verify the frequency of the terms used in specific periods, thereby assisting in the historical analysis and identification of possible trends. The classification of information in InfoSIC is divided into three levels.

*Classification level 1 - links in the production chain*

- Components · Leather · Artifacts · Machinery · Retail · Shoes.

*Classification level 2 – categories*

The categories, also called “Walk the Factory Floor” classification, serve to classify and organize the contents in InfoSIC as well as in the InfoSIC Portal – an important difference since in InfoSIC, the classification operations environment, the handling and management of the terms are determined by the experts. In the SIC portal the categories are responsible for relating common themes, making them available in one location, besides the production chain (another item also used for classification). The categories were designed so as to be able to expand them according to the classification needs of new themes within the supply
chain, under potential expansion. The expressions, considered as regular in occurrence within this context, are listed below:

<table>
<thead>
<tr>
<th>Foreign Trade</th>
<th>Institutions and Associations</th>
<th>Productive Poles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economy</td>
<td>Legislation and Taxation</td>
<td>Production</td>
</tr>
<tr>
<td>Employment</td>
<td>Marketing</td>
<td>Product</td>
</tr>
<tr>
<td>Company</td>
<td>Environment</td>
<td>Intellectual Property</td>
</tr>
<tr>
<td>Events</td>
<td>Fashion</td>
<td>Quality</td>
</tr>
<tr>
<td>Financing and Credit</td>
<td>Marketing</td>
<td>Human Resources</td>
</tr>
<tr>
<td>Business Management</td>
<td>Environment</td>
<td>Service</td>
</tr>
<tr>
<td>Innovation</td>
<td>Marketing</td>
<td>Technology</td>
</tr>
<tr>
<td></td>
<td>Environment</td>
<td>Sales</td>
</tr>
</tbody>
</table>

*Classification level 3 – keywords*

For the registration data entered into the system, the key concepts that represent the specific content of such data will be extracted. Within the InfoSIC system, these concepts are considered keywords that represent the subjects. These are elements considered the connecting axes of the contents in the system that gather contents on the same subject in the portal and that also help to search for information.

The terminology list proposed to guide and control the vocabulary of the leather, footwear and artifacts industry, in InfoSIC, was initially categorized based on the divisions suggested by the Classification Research Group (CRG), and adapted to the context of the sector (Table 1) by the authors of this work. Therefore, the key words selected to represent the contents in the system should be controlled by the Terminology List. However, these words can be related to the general categorizations: Links in the chain and the Factory Floor. Below is the adaptation of the CRG categories for the footwear sector:

**Table 1: Categories suggested by CRG adapted to the Footwear Industry**

<table>
<thead>
<tr>
<th>Classification Research Group (CRG)</th>
<th>Categories leather / footwear</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miscellaneous, substances, entities.</td>
<td>Products: shoes, tennis shoes, boots</td>
<td>Articles produced in the footwear chain.</td>
</tr>
<tr>
<td>Products, Instruments</td>
<td>Equipment (machinery, tools and instruments): e.g., hammer</td>
<td>All equipment used in the production of footwear</td>
</tr>
<tr>
<td>Mental constructs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part Constituents</td>
<td>Material: leather, rubber</td>
<td>Base material used in the manufacture of footwear.</td>
</tr>
<tr>
<td>Section</td>
<td>Elements: heels, Luis XV, Anabela, etc.</td>
<td>Parts of the shoes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Systems</td>
<td>Not yet unidentified</td>
<td>Not yet unidentified</td>
</tr>
<tr>
<td>Overall attributes, Qualities, properties, measures</td>
<td>Models: clogs, pumps</td>
<td>Models of the product.</td>
</tr>
<tr>
<td>Structure, Process, behavior</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Action object (recipient)</td>
<td>Occupations: Backstitch seamster</td>
<td>Names of occupations that make up the footwear production chain process.</td>
</tr>
<tr>
<td>Relationships, interaction.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reactions</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Classification Research Group (CRG)

<table>
<thead>
<tr>
<th>Categories leather / footwear</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall operations</td>
<td>Processes: Modeling, retanning, etc.</td>
</tr>
<tr>
<td>Experiments, tests</td>
<td>All processes, methods and steps in the footwear manufacture.</td>
</tr>
<tr>
<td>Mental operations</td>
<td></td>
</tr>
<tr>
<td>Properties of attributes,</td>
<td>Property: e.g. dyeing affinity</td>
</tr>
<tr>
<td>relationships and operations</td>
<td>The characteristics, properties and qualities of products, components.</td>
</tr>
<tr>
<td>Location, condition</td>
<td>Location: e.g., shoe shop, company, association</td>
</tr>
<tr>
<td></td>
<td>Geographic areas and environments (manufacture, sales, repairs, etc.)</td>
</tr>
<tr>
<td>Time</td>
<td>Dates</td>
</tr>
<tr>
<td></td>
<td>The periods, time of year, seasons</td>
</tr>
</tbody>
</table>

In the system interface (Figure 1) these categories were systematized as follows:

**Fig. 1: Interface of InfoSIC with the thematic categories for terminology management**

The terms and their hierarchical relationships and equivalence are within these categories. The relationships to determine the preferred terms to be used in the system are set out in the InfoSIC interface.

For the initial survey of terms, to compile this terminology list, the elaboration of a balanced *corpus* was prepared by analyzing the texts of the area, reading the texts
(permanent activity, therefore, not described in this paper); use of software for word mining (E-terms), organization of terms in the categories, classes and subclasses, selection of the terminology and organization of lexical units, validation of terms and its structure by the specialist in the area, adjustments in the Terminology List according to the suggestions by the specialist (Trybulia, 1999). The corpus produced consisted of 709 texts that included various regions of Brazil, of which 509 were news related articles, 123 were articles, 77 were dissertations and 11 were blogs.

Besides the Auto-Search feature, all terms selected from the specialized documents, perceptions and user searches were organized within the categories, which are structured and systematized in InfoSIC according to hierarchical distribution.

With the mechanism conceived, the concepts selected from the reading and documentary analysis should be searched in this hierarchical list in order to be validated as “descriptors”, keywords of the registered contents in InfoSIC.

In these methodologies we introduce space structures to fit a candidate term. This term will then be organized by the librarian in charge of this task. This candidate term can be suggested to the classifier during the indexing process or can be suggested to the system user when searching for information. The addition of these terms has provided robustness to the system given that new concepts continuously appear, especially terms related to fashion. And in the economic market, fashion promotes the footwear production and consumes.

**Fig. 2: Interface of InfoSIC with details of the Terms selected to represent a specific content**
We understand that this methodology can promote information retrieval that is imprecise to some degree. An important aspect in the information organization process in this system is that the information will be indexed using the terminology list, however, the information will also be classified into 6 categories related to footwear production and will be simultaneously classified into 22 categories related to a footwear production process. The chain connection represents the production sectors involved in the footwear domain and it was suggested by footwear professionals.

We were very cautious regarding the building of the controlled vocabulary when using the contents published regarding the sector. Specialists were consulted the specialist. The potential users were consulted when terms used for search were selected. Quantitative guaranties, when we selected the terms that appeared frequently whenever the E-terms were used. After organizing the information products, an information analysis professional is selected and an abstract is prepared after the subject and terms to be classified and indexed are selected.

The subject categorization in a terminological list is vital to organize the domain; however, it is not a rule to use the same real categorization to represent a specific domain. This helps the indexer to select and systematize the terms in a language. The categorization in our list does not need to represent the domain practices. The relationship of the controlled vocabulary with the subject domain will be by cross referencing the Chain Connection classifications and then the Production Step is carried out. For example: If in the system we register news about: „Use of acrylic shoe for the 2012 summer station”; first we select and download news in the system, develop the documentary analysis and then select the terms used in the text of the news. Next, we index the news in a specific field. In the terminology list, the term acrylic will be organized in the category as: our constituent’s parts – materials – elements. We will then add this definition: Acrylic is a type of clot or plastic produced by chemical process. Note: All the words used in this definition can be used with a keyword to search the news. Next, there is the equivalent relationship with their synonyms. When acrylic is used to represent news in the system, this term can be classified in more than one Chain Connection (i.e. Fashion) and Production Step (Production). These allow the term acrylic to be related in different contexts independent of their location in the terminology list. Using these specifications, the entire subject can be indexed and retrieved according to the nature of the domain.

Final Thoughts

Infosic is still in the adjustment and consolidation stage, both as regards its operating structure as well as its software structure, and also its own methodology, produced to meet the demand identified in the information classification process in this field. However the information that was presented in this work foresees the possibility of using the language instruments as tools for competitive intelligence within a production sector, in particular the footwear manufacturer.

The goal of this paper was to demonstrate how a three level cross-reference classification methodology – which categorically includes the information production structure of the sector – can be a strategy that promotes relevant results in information retrieval.

To conclude, we state that the main objective of this study was to reinforce the fact that there is an increasing demand by the manufacturing sectors for systems and methodologies that facilitate assembling the organization, classification and retrieval of information with added value.
References
Information Mining and Visualization of Data from the Brazilian Supreme Court (STF): A Case Study

Abstract
This paper describes a joint research of the Law School (Direito Rio) and the Applied Math School (EMAp) of the Getulio Vargas Foundation (FGV), Brazil to analyze information from judicial activities in some of the Brazilian courts. The data for the study included the entire collection of judicial decisions from 1988 to the present. The idea was to identify bottlenecks in the judicial processes at the STF.

Introduction
Large collections of textual data present a substantial challenge for extraction of relevant bits of information to feed subsequent statistical analysis and visualization pipelines. The peculiarities of the knowledge domain often require the implementation of customized natural language processing pipelines, along with specific knowledge organization systems, to describe the relevant terminology. This paper describes a joint research made by the Law School (Direito Rio) and the Applied Math School (EMAp) of the Getulio Vargas Foundation (FGV), Brazil. After initial contacts a joint venture was established between researchers from EMAp and Direito Rio to analyze the information from judicial activities, in some of the Brazilian courts. Initially, the Law School intended to analyze the behavior of the Brazilian Supreme Court (STF) to support public policy-making, and to identify bottlenecks in the judicial processes at the STF. The task was to analyze the texts of the entire set of recorded judicial decisions, accessible through the STF institutional site.

This data had never been analyzed on this scale before, so a great deal of exploratory analyses was expected in order to reveal hidden patterns in the data. A number of \textit{a priori} questions were proposed, basically aimed at determining if the STF was performing according to its constitutionally defined role, and if not, in what way it could be changed to better serve its purpose.

Some of the methodology described herein was applied to the generation of the results published in the project’s first technical report (“I Relatório – abril/2011 – O Múltiplo Supremo”, 2011). The results presented here go more in the general direction of exploratory data analysis.

Methodology

\textit{Data Retrieval}

The data for the analysis consisted of the entire collection of judicial decisions from 1988 to the present. This data although open to all, is available to the public only through a simple web-based query interface. Given the dimension of the dataset manual retrieval was out of the question. Therefore, a software tool was developed to retrieve the documents. This
process resulted in a collection of approximately 1.2 million cases (HTML documents). Due to the fragility of the STF web servers, the data retrieval phase lasted approximately four months.

In order to facilitate information extraction, the original documents, in HTML format were parsed and stored on a relational database for further analysis. We sought to preserve as much structure as possible when analyzing the documents retrieved. Figure 1 has an example of the kind of relational structure we manage to elicit from the raw documents. From these initial tables a large number of tables were derived for purposes of analysis.

**Fig. 1: Main tables of the resulting relational database**

The next big challenge was to extract specific information from the large collection of texts in “t_decisoes.decisao”, containing the full text of each decision.
**Extracting Information**

Most of the time, it is not difficult to deal with quantitative aspects of information extracted from documents. Identifying specific words, their frequencies and collocations are the basis of quantitative textual analysis. Techniques like identifying Term frequency / Inverse Document frequency and the extraction of noun phrases allowed us to identify the subject of the cases and their most used expressions. We were also able to identify the main litigants, the evolution of the legal cases through the last decades, the lifespan of each case and so on. Data such as names, dates, locations, legal citations, case subjects, etc. were all analyzed, in order to characterize STF processes and the judges’ work. After identifying the law citation signature - using a regular expression - we were able to associate the set of laws that were cited in each legal case (Fig. 2).

**Fig. 2:** Example of usage of regular expression to extract patterns of law citation from the decision text in HTML format

```text
>.*s*([A-Z]{3}-[A-Z,0-9]{3,6})\s+
<pre>LEG-FED LEI-005869 ANO-1973 ART-00267 INC-00006
CPC-1973 CODIGO DE PROCESSO CIVIL
LEG-FED DEL-002665 ANO-1987 ART-00008
LEG-FED DEL-002666 ANO-1987 ART-00001
LEG-FED SUM-000266
(STF).</pre>
```

From the sets of citations extracted from each decision, we proposed some metric for law usage based on Shannon's entropy (Bommarito II & Katz, 2010). The rationale behind it was that the entropy would be greater for cases in which the decision was based on wider jurisprudence while decision based on a single citation to a law, would give zero entropy.

However, for the quintessential information to be mined, it was necessary to resort to more complex natural language processing (NLP) techniques to try to elicit and „understand” the „meaning” of some portions of the texts. Again, the development of software for NLP and a small set of STF specific, lightweight domain ontologies (Fig. 3) we have built, allowed us to take further steps on the analysis of the material.

**Information Visualization**

A good strategy to identify hidden patterns in data is through visualization. Our tasks involved building graphs (networks) to connect various bits of information extracted from the texts, depicting the established relationships. Due to the size and complexity of our dataset, graphs with hundreds of thousand (and sometimes many more) nodes were common. The graphs allowed us to identify, for example, clusters of judges that were employing similar sets of laws, and animations permitted us to see how laws were used over time. These graphs, though, not shown here due to their complexity were instrumental in the identification of many patterns which led to other analyses.

**Results**

The application of the light ontology depicted on Figure 3 allowed us to classify and analyze the stages, for example, with respect to duration (Fig. 4). This dissection of the cases allowed us to pinpoint stages which contributed more to the total lifespan of a judicial case. Another important aspect revealed, was the multi-modal nature of the duration distributions, which indicates some severe bottlenecks for certain types of cases. Once identified, these bottlenecks can be further investigated and addressed.
Fig. 3: Light Domain Ontology of the stages of the life history a legal case. Names are maintained in the original language as their translation would convey little meaning. This diagram is here to illustrate the complexity of the taxonomy required to understand the data.
Fig. 4: Distribution of durations of the "reautuação" stage. The left y-axis is the absolute number of this kind of stage in the database and the right y-axis is the normalize scale of the density plot (in green).

Fig. 5: Entropy increase with Time. This effect may be due to the improvement in the accessibility of jurisprudence with time. For this graph the entropy was calculated on the citations aggregated over a year.
The entropy of decisions, which was proposed as a proxy to the complexity of decisions as measured by the jurisprudence applied, showed a marked increase in time (figure 5). This may reflect the technological improvements in the accessibility of relevant jurisprudence.

Conclusions
This paper reports a pioneering piece of work on large scale analysis of the dynamics of the Brazilian supreme court. The project's initial technical report (“I Relatório – abril/2011 – O Múltiplo Supremo”, 2011), reported important distortion in the work load of STF and has led the president of the STF, to propose a constitutional amendment to fix these distortions.

One of the biggest challenges we have identified in this dataset, is the difficulty in asserting the outcome of a decision (Kastellec, 2010). The cryptic and vague language used in the verdict makes it very hard even with the help of advanced NLP techniques, to know which part the decision favors. Additionally, the frequent multiple appeals complicate matters even further.

The richness of the data is unquestionable, and to explore it we required advance mathematical and computational techniques to extract, analyze and visualize information. The initial stages of this work have taken advantage of some well established analytical tools which, nevertheless had to be adapted to work with large volumes of data. Distributed processing versions of these tools are particularly lacking, leading our group to invest in the development of such tools. These developments take time, but paint a bright future for the analyses of large text corpora.

In this paper, we sought to present some of the challenges in the analysis of the entire recorded history of the Brazilian Supreme Court. We hope that the preliminary analyses presented in this paper, are sufficient to illustrate the richness of this dataset, and its potential for further studies. This is an ongoing project, and the tools and insights developed are already bearing fruit to other analytical projects targeting other Brazilian courts, much larger in size.

References
Carlos Alberto Corrêa - Universidade de São Paulo (USP), Brazil
Nair Yumiko Kobashi - Universidade de São Paulo (USP), Brazil

Automatic Indexing and Information Visualization: A Study Based on Paraconsistent Logic

Abstract
This paper reports a research to evaluate the potential and the effects of use of annotated Paraconsistent logic in automatic indexing. This logic attempts to deal with contradictions, concerned with studying and developing inconsistency-tolerant systems of logic. This logic, being flexible and containing logical states that go beyond the dichotomies yes and no, permits to advance the hypothesis that the results of indexing could be better than those obtained by traditional methods. Interactions between different disciplines, as information retrieval, automatic indexing, information visualization, and nonclassical logics were considered in this research. From the methodological point of view, an algorithm for treatment of uncertainty and imprecision, developed under the Paraconsistent logic, was used to modify the values of the weights assigned to indexing terms of the text collections. The tests were performed on an information visualization system named Projection Explorer (PEx), created at Institute of Mathematics and Computer Science (ICMC – USP São Carlos), with available source code. PEx uses traditional vector space model to represent documents of a collection. The results were evaluated by criteria built in the information visualization system itself, and demonstrated measurable gains in the quality of the displays, confirming the hypothesis that the use of the para-analyser under the conditions of the experiment has the ability to generate more effective clusters of similar documents. This is a point that draws attention, since the constitution of more significant clusters can be used to enhance information indexing and retrieval. It can be argued that the adoption of non-dichotomous (non-exclusive) parameters provides new possibilities to relate similar information.

Introduction
Information retrieval systems are traditional environment for testing automatic indexing quality. However, other systems also depend on these solutions, such as information visualization systems.

Automatic indexing and information visualization involves many disciplines and theories. This paper presents the results of a study on the effects of the use of Paraconsistent logic in automatic indexing, using a visualization system as a test environment. The work is organized into the following sections: Automatic indexing and Vector Space Model; Automatic indexing and non-classical logic; Automatic indexing and Paraconsistent logic; Visualization, similarity, distance and proximity; Computational experiment; Final considerations.

Automatic Indexing and the Vector Space Model
One of the most used and cited models of automatic indexing and visualization is the vector space model (VSM), developed by Luhn and popularized by Salton in SMART – System for the Manipulation and Retrieval of Text (Raghavan, 1997).

The vector approach is based on the mathematical concept of vector. Salton and McGill (1983) consider that the distance between two vectors representing documents indicates the degree of similarity between them. There are several formulae to calculate distance. One of the most used is (Salton; McGill, 1983):

\[ sm(x, y) = \frac{\sum_{i=1}^{n} (w_{ix} \times w_{iy})}{\sqrt{\sum_{i=1}^{n} (w_{ix})^2} \times \sqrt{\sum_{i=1}^{n} (w_{iy})^2}} \]

where \( w_{ix} \) refers to the weight of \( i \)-th term (or element) of vector \( x \), and \( w_{iy} \) refers to the weight of \( i \)-th word (or element) of the vector \( y \). Once understood the basics of VSM, one question remains: how to set values to the weights of indexing and searching terms or descriptors?
Various methods are cited in the literature (Salton; McGill, 1983; Salton; Buckley, 1988). The most widely used is based on the statistical frequency of occurrence of terms in two different instances: in the document individually, and in the collection. In the first case, the value is called $tf$ (term frequency). In the second, the measure is called $idf$ (inverse document frequency), calculated by $idf_t = N / n_t$, where $N$ is the total number of documents of the collection and $n_t$ is the number of documents that contains at least one occurrence of the term $t$. A variant of this formula uses the base-10 logarithm, so $idf_t = \log(N / n_t)$. Once defined the values of local and global instances of term $t$, its weight is calculated as follows: $W_{t,d} = tf_{t,d} \times idf_t$.

**Automatic Indexing and Non-Classical Logic**

The solutions developed in automatic indexing often use theories such as non-classical logics to deal with uncertain, imprecise, vague and ambiguous situations. Two examples of non-classical logics are Fuzzy logic and Paraconsistent logic. Fuzzy logic was developed based on the theory of fuzzy sets proposed by Loft Zadeh (Bojadziev; Bojadziev, 1995). Paraconsistent logic was developed, independently, by Stanislaw Jaskowski (Poland) and Newton da Costa (Brazil). Its main features are: (i) suppression of the principle of non-contradiction, (ii) acceptance of various logic states beyond True and False.

**Automatic Indexing and Fuzzy Logic**

Some researchers believe that Fuzzy logic can be used to deal with vagueness and subjectivity in indexing procedures, as well as to manage the vagueness embedded in queries (Viedma-Herrera, Pasi, 2003).

Bordogna and Pasi (1995) applied Fuzzy logic in information retrieval taking into account that: (i) the documents are organized into structural parts, such as title, author(s), keywords, abstracts and references. Thus, the informational role of each term depends on the section in which it occurs, (ii) sections of a document may have different degrees of importance for each user and thus to calculate the degree of significance depends on the intervention of the user.

Another approach of Fuzzy logic was made by Molinari and Pasi (1996). They proposed the use of the syntactic structure of language and text for indexing HTML documents. The model assigns to each section a degree of importance. Thus, the characteristics of HTML documents, such as italic, bold and font size, receive different values, according to pre-defined criteria.

**Automatic Indexing and Paraconsistent Logics**

The use of Fuzzy logic in automatic indexing encouraged the use of Paraconsistent logic. This study adopts one of its variants - the Paraconsistent Annotated Logic with annotation of two values (PAL2v). According to this logic, two variables are assigned to a proposition: the degrees of belief and unbelief, with values ranging in the interval $[0, 1]$. This approach allows establishing four main logical states (inconsistent, true, false, indeterminate), and a variable amount of intermediate logical states. The logical states are determined from the analysis performed on two annotated values, described by a pair $(\mu_1, \mu_2)$, which represents the degrees of belief and unbelief assigned to a proposition. Thus, for a given proposition, if the values of the degrees of belief and unbelief are established or calculated as the values $(1.0, 0.0)$, the meaning will be: total belief and lack of unbelief in the proposition. Their logical state is true. Similarly one can describe other three main logical states: $(1.0,1.0)$ – inconsistent (total belief and total unbelief); $(0.0,1.0)$ - false (lack of belief and total unbelief), $(0.0,0.0)$ – indeterminate (lack of belief and lack of inbelief). Other values within the interval $[0,1]$ can be attributed to degrees of belief and unbelief, in order to establish intermediate logical states.

The intermediate and main logical states of the PAL2v could be represented on a graph in which the horizontal and vertical axes indicate the logical degrees. The graph can represent
logical states and regions associated with them. Thus one can, from the numerical values of
the factors of belief and unbelief, establish the logical state associated with them. This
graph is called Unitary Square of the Cartesian Plane (USCP).

The Fig. 1 shows this configuration. One can observe three types of features: (i) the main
logical states: True (1.0,0.0), Inconsistent (1.0,1.0), False (0.0,1.0) and Indeterminate
(0.0,0.0); (ii) some intermediate logical states: False tending to Inconsistent (1/2,1.0),
Almost true (1/2,1/2), False tending to Indeterminate (0.0,1/2), True tending to be
Inconsistent (1,1/2), False tending to Indeterminate (0.1/2) and True tending to
Indeterminate (1/2,0.0); (iii) regions associated with both the main logical states and the
intermediate logical states: T - Inconsistent, F - False; ⊥ - Indeterminate: V - True, ⊥ → f -
Indeterminate, tending to False; ⊥ → V - Indeterminate, tending to True, T → f -
Inconsistent, tending to False; T → v - Inconsistent, tending to True, Qv → T – Almost
tending to Inconsistent, Qf → T – Almost False, tending to Inconsistent, Qf → ⊥ -
Almost False tending to Indeterminate; Qv → ⊥ - Almost True, tending to Indeterminate.

**Fig. 1: USCP of resolution 12 - highlighting intermediate and extreme regions**

Source: Da Costa et al. (1999), p. 89

Da Costa et al. (1999) describe other variables that may be used to delimit the regions.
They are: \( V_{sc} \) - superior value of certainty control, which limits the degree of certainty
around the true region; \( V_{ic} \) - lower value of certainty control, which limits the degree of
certainty around the false region; \( V_{scct} \) - superior value of contradiction control, that limits
the degree of contradiction near the inconsistent state; \( V_{ect} \) - lower value of contradiction
control, that limits the degree of contradiction close to indeterminate region. For the
previous graphic the variables assume the following values: \( V_{sc} = 1/2; \) \( V_{ic} = -1/2; \)
\( V_{scct} = 1/2; \) \( V_{ect} = -1/2. \)

The variables are not directly visible in the graphs. They are mathematical artifices to
adjust the size of the regions defined in USCP. With these adjustments it is possible to
calibrate so as to restrict the sensibility to certain logical values. Fig. 2 is a variation of
Figure 1, in which the values of variables \( V_{sc}, V_{ic}, V_{scct} \) and \( V_{ect} \) were modified.

In Fig 2, the values are: \( V_{sc} = 3/4, \) \( V_{ic} = -3/4, \) \( V_{scct} = 1/2 \) and \( V_{ect} = -1/2. \) One can see
the reduction in size of the regions of True and False. In this case, an analysis considering
the values of belief and unbelief (\( μ_1 \) and \( μ_2 \)) will be more critical for such logical state. That
is, the number of situations in which the pair \( μ_1 - μ_2 \) will be considered True or False is
smaller than the number of situations shown in Fig. 1.

**Fig. 2: USPC of resolution 12 - highlight for regions with modified size**

Source: Da Costa et al. (1999), p. 91
Although the figures are good tools to visualize the logical state of a proposition, they are not useful to assessments on an ongoing or repetitive way. Thus, Da Costa et al. (1999) developed an algorithm that allows, from the values of the degrees of belief and unbelief, to determine the logical state of a proposition from the values of $\mu_1, \mu_2, \nu_{sc}, \nu_{cc}, \nu_{sect}$ and $\nu_{cct}$.

**Visualization, Similarity, Distance and Proximity**

Computer Graphics aims to promote human-computer interaction and computer manipulation of images. These studies are considered potentially useful to minimize the problems of noise or silence in Information Retrieval. The solutions allow users to know better the search space in which they are working, thus increasing effectiveness. These solutions can reveal the semantic structure of a collection of documents, allowing one to use this view as the basis for subsequent searches (Borner; Chen; Boyack, 2003).

This Projection Explorer (PEx), developed at the Institute of Mathematics and Computer Science - Universidade de São Paulo (Campus of São Carlos) is a system to create document collection visualizations using small colored circles to represent them (Paulovich; Oliveira; Minghim, 2007; Paulovich et al., 2008).

Systems developed to display document collections perform calculations and procedures with two specific objectives: (i) the grouping of similar documents, (ii) the separation of dissimilar groups of documents. For this, the systems must be able to indicate the degree of similarity between the components of each group of documents. One of the models used for this purpose is the VSM.

Fig. 3 is an example of visualization performed with PEx. The circles with the same color indicate the documents of the same subject (this figure was originally created with four different colors), they all belong to the same test collection. In this Figure, the collection has four subjects. The rectangles highlight groups of documents which, although of different subjects appear close. This effect is a consequence of the system criteria to build the Figure. In this case, the documents present some degree of similarity, though from different subjects. In this context, it may be asked whether such results could be easily observed in an information retrieval system.

![Fig. 3: Visualization of a document collection](http://infoserver.lcad.icmc.usp.br/infovis2/PEx)
Computational Experiments
The tests performed in this research involved the para-analyzer algorithm in PEx system. The tests can be summarized in the following steps: (1) generating the visualization of a collection of documents using the calculation facilities of the indexing/similarity originally built in PEx; (2) use of the para-analyzer algorithm to recalculate the weights established for the documents of the collection; (3) getting a preview of the collection with these new weights; (4) comparing the view obtained in (3) with that obtained in (1).

Test Collections
The main test collection is called CBR-IR-ILP-SON, composed of the initials of the subjects of the documents: CBR - Case-based reasoning; IR – Information Retrieval, ILP – Inductive logic programing; and SON – Sonification. The collection had 675 documents. According to Paulovich (2008), the documents of CBR and ILP were retrieved from scientific journals. IR documents and SON documents were obtained by searching the internet and ranked in an approximate way, according to the source from which they were obtained (Minghim; Levkovits, 2007).

New collections were built from the original collection: CBR-IR-ILP, with 574 documents; CPR-ILP-SON, with 496 documents, CBR-ILP, with 395 documents and IR-SON with 280 documents. The collections, in the system, were built from the titles, abstracts and references for each document.

Evaluation of Visualizations
The PEx system has facilities for evaluating the quality of views, taking into account different aspects.

One such method is the use of a value to measure the quality of the clusters produced by different algorithms. These methods are based on Cluster Analysis. One of these measures, called silhouette coefficient (Rosseeuw and Kaufman, 1990); (Tan, Steibach and Kumar, 2006) is a value between -1 (worst value) and 1 (best value). Kaufman and Rosseeuw (1990) produced a table in which the quality of a group can be evaluated according to the coefficient value obtained for it. Thus, for values in the interval [0.71, 1.00] it can be stated that a strong structure was found; for values in the range [0.51, 0.70] a reasonable structure was found; for values in the range [0.26 , 0.50] the structure is weak and can be artificial, and for that set, additional methods have to be tried; for values lower than or equal to 0.25, no substantial structure was found.

The use of Para-analyzer in Projection Explorer
The test was carried out in three stages. First, we included the para-analyzer algorithm in PEx, allowing a different way to calculate the weights used by VSM. Second, we chose one proposition: the term t is a good term to index the document D. As the para-analyzer algorithm is capable of evaluating the logical state of a proposition, it will be possible to evaluate the quality of the term using that algorithm. Third, it was assigned a penalization to the original weight (calculated using the tf-idf approach) of the term (or its degree of utility), according to the region indicated by the para-analysis. For this procedure it was chosen the value tf of the term as the degree of belief and the value df (the inverse of idf) as the degree of unbelief. The originally calculated weights were kept only for the terms whose para-analysis indicate the region of true (so, for this region the factor is 1). For the other regions, the penalization factors were defined considering a decrease of 0.15 of value 1, as the region move away from the region of true. So, suppose that the calculation of the weight of a term using the formula tf * idf is equal to 100 and, on the other hand, if the para-analysis of the term indicates the region Qv → ⊥ (almost true tending to indeterminate), and as this region has a penalty factor of 0.85, the new weight to be given to the term shall be 0.85 * 100, or 85. The aim is to observe the effects that the change of the weight of the term will cause in the collections display.
In addition to the penalty of the terms, the tests were performed modifying the sizes of the USCP regions. This was made by manipulating the variables $V_{scc}$, $V_{icc}$, $V_{scct}$ and $V_{icct}$ as described above. The regions used in the tests and their limits are: Region 0 (R0) $V_{scc} = 1/2$, $V_{icc} = -1/2$, $V_{scct} = 3/4$, $V_{icct} = -3/4$; Region 1 (R1): $V_{scc} = 1/2$, $V_{icc} = -3/4$, $V_{scct} = 3/4$; Region 2 (R2) $V_{scc} = 3/4$, $V_{icc} = -3/4$, $V_{scct} = 3/4$, $V_{icct} = -3/4$; Region 3 (R3) $V_{scc} = 1/2$, $V_{icc} = -1/2$, $V_{scct} = 1/2$, $V_{icct} = -1/2$.

The procedure described above produced different layouts from those produced with the original facility embedded in PEx.

**Results**

The results were compiled for each collection. In this paper we present the values of the silhouette coefficient of each visualization. The curves are identified by the names of the collections, and Rx is used for the views obtained by using the para-analyzer in the region Rx, where x can be 0, 1, 2 or 3 as described above.

The results were obtained checking the values of the silhouette coefficients for the constructed visualizations. The expression "original" (Table 1) refers to the figures obtained with the original procedure of PEx. The results are presented in table and graph:

<table>
<thead>
<tr>
<th>Collection / Region</th>
<th>Original</th>
<th>R0</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBR-ILP-IR-SON</td>
<td>0.05844</td>
<td>0.18286</td>
<td>0.28807</td>
<td>0.20883</td>
<td>-0.04546633</td>
</tr>
<tr>
<td>CBR-ILP-IR</td>
<td>0.19003</td>
<td>0.23321</td>
<td>0.32509</td>
<td>0.65129</td>
<td>0.22175422</td>
</tr>
<tr>
<td>CBR-ILP-SON</td>
<td>0.41562</td>
<td>0.41507</td>
<td>0.57681</td>
<td>0.51775</td>
<td>0.66061221</td>
</tr>
<tr>
<td>CBR-ILP-IR-SON</td>
<td>0.67798</td>
<td>0.66479</td>
<td>0.48078</td>
<td>0.74633</td>
<td>0.010869439</td>
</tr>
<tr>
<td>IR-SON</td>
<td>0.4099</td>
<td>0.40576</td>
<td>0.39995</td>
<td>0.54973</td>
<td>-0.040663127</td>
</tr>
</tbody>
</table>

*Source: obtained with Projection Explorer*

**Graph 1: Comparison of the silhouette coefficients**

The procedure of changing the weights of indexing terms using the para-analyzer improved the silhouette coefficient in certain regions. Because this coefficient is used to evaluate the quality of the obtained clusters and as no other change was made in PEx parameters, it can be concluded that under certain conditions the modification of weights improved the quality of the clusters.

**Final Considerations**

This research required interactions among different disciplines: information retrieval, automatic indexing, non-classical logic and information visualization. The research was developed through two perspectives: 1) modifying the traditional calculation of weights used in VSM, using methods based in Paraconsistent logic; 2) measuring the effects in the
calculation of weights in a visualization system - the Projection Explorer (PEx), a system based on VSM.

The effects produced in the clusters indicate that the use of para-analyzer, under the conditions of the experiment, has the capacity to generate more effective clusters. The use of para-analyzer produced measureable gains in the evaluation of the clusters that represent the visualizations, using the silhouette coefficient.

The computational experiments carried out showed the potential of the para-analyzer in automatic indexing. However, it was not able to explain why and how the observed effects occurred. It can be argued, however, that the adoption of non-dichotomous parameters (non-exclusive) provides new possibilities to link similar information. Other experiments must be carried out in order to broaden the understanding of the detected effects.

References
Digitized Content and Index Pages as Alternative Subject Access Fields

Abstract

This article describes a pilot study undertaken to test the benefits of the digitized Content and Index pages of books and content pages of journal Issues in providing subject access to documents in a collection. A partial digitization strategy is used to fossick specific information using the alternative subject access fields in bibliographic records. A pilot study was carried out to search for books and journal articles containing information on “Leadership”, “Women Entrepreneurs”, “Disinvestment” and “Digital preservation” through normal procedure and based on information stored in MARC 21 fields 653, 505 and 520 of the bibliographic records in the University of Mumbai Library. The results are compared to draw the conclusions.

Introduction

Keyword searches in OPACs usually begin by looking for the presence of the keyword in the title of a document. The searcher may then use the output as a guideline to refine the search based on a perusal of subjects headings assigned to the initially retrieved documents. Given the fact that title of every document does not necessarily include the search terms, the searcher may also want to browse the content and index pages of documents that are not listed in either title or subject search. Users depend on primary journals for current information. Unless a library has an in-house indexing and abstracting service / table of contents service the user identifies relevant papers by browsing tables of contents of journals. Given this situation, it was thought that it is useful to make certain portions such as content page, preface/introduction, index pages of books and abstracts / introduction / objectives / observation / conclusion portions of a journal article searchable in an OPAC.

Objective

The objective of this paper is to compare results of subject searches using traditional bibliographic data and free text data extracted from select digitized portions of documents. This is achieved by compiling the relevance ratio of number of hits using different search results to the total number of all books in a class number which represents the subject sought:

a. Search result before digitizing uncontrolled terms and title + subject;
b. Search result after digitizing uncontrolled terms (as SOUL 2.0 includes uncontrolled terms in subject search);
c. Compare the results, search procedures to suggest additional data entry module options and search options of the OPAC module to enhance the search results.

Methodology

The Library of the University of Mumbai, which is one of the large academic libraries in India, has a collection that exceeds eight hundred thousand documents. The library uses the SOUL software for library automation. The Jamnalal Bajaj Institute of Management (JBIMS) is one department of the University. This experiment was carried out in the library of this institute and this paper reports a pilot study undertaken to test the benefits of the proposed partial digitization strategy for capturing portions of texts to serve as an alternative subject access field in the bibliographic database. In the pilot study, books and journals containing information on “Leadership”, “Women Entrepreneur/s”; “Disinvestment Divestment”
and “Digital Preservation” are first searched in the bibliographic database without any digitized text (content and index pages). “Digital Preservation” is first searched through the free text search option of SOUL 1.0 OPAC module of university library. Remaining search terms i.e. “Leadership”, “Women Entrepreneur/s”, “Disinvestment/Divestment” are searched by title, title + subject using the simple search option of SOUL 2.0 OPAC module. From this search results the class number is identified for each term under which books on the subject are classified. A bibliography is compiled and each book’s content and index pages are checked to get the sought information.

Now the bibliographic database is edited to add the digitized content and index pages to the „Uncontrolled terms” field - MARC-21 Tag 653. All the terms are once again searched through subject in OPAC and results are noted in Table1.

For searching journals, files containing photocopied content pages were browsed manually and synonym/s as well as acronym/s was/were identified for the search terms. These were noted for compiling the thesaurus and master file of acronyms. Whereas for digital preservation the synonyms were not identified, but some broader terms like digital library/libraries were used. The University of Mumbai Library has been maintaining content pages file since 1997 for journals in library science as a pilot project; and at JBIMS this is being done from 2010.

The same is then achieved after digitizing the content pages of these journals and stored in „Formatted Content” MARC 21 Tag 505 of SOUL 2.0 „Article indexing” module. The results are noted in Table1.

### Table 1: Search results of bibliographic database, bibliographic + digitized text and respective relevance ratios

<table>
<thead>
<tr>
<th>Row No.</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Primary Search terms</td>
<td>Leadership</td>
<td>Women Entrepreneurs</td>
<td>Disinvestment/Divestment</td>
<td>Digital preservation</td>
</tr>
<tr>
<td>2</td>
<td>Synonym/s &amp; acronym/s for the search terms</td>
<td>CEO, CFO; Success/succeeding/succession &amp; Managers; Top employees; High potential managers; Executive/s; Leaders; Managers</td>
<td>women finance leaders; Women in Management; +Digital Library/libraries;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Title search result</td>
<td>38</td>
<td>2</td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>4</td>
<td>Title + Subject Search Result</td>
<td>62</td>
<td>2</td>
<td>2</td>
<td>37</td>
</tr>
<tr>
<td>5</td>
<td>class number search result</td>
<td>353</td>
<td>10</td>
<td>66</td>
<td>79</td>
</tr>
<tr>
<td>6</td>
<td>content Page search result</td>
<td>*</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>Index Page search result</td>
<td>*</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>Search result for title+subject after digitizing the content and index pages</td>
<td>*</td>
<td>5</td>
<td>4</td>
<td>43</td>
</tr>
</tbody>
</table>
To speed up this process and for precision, content and index pages of these books and content pages of journal issues are scanned and saved as image files. These images are then converted into optically recognized characters.

**Observations**

1. It is observed that from the manual search of content pages there were various terms under which articles for “Leadership”, “Women Entrepreneur/s”, “Disinvestment/Divestment”, and “Digital Preservation” can be searched for. The Rows (B/C/D/E,2) lists these relevant terms under which sought information can be looked for;

2. Row (B/C/D/E,3 to B/C/D/E,7) gives the number of hits(books) for the title; the title + subject; the class number; the content page and the index page search under each respective search terms. The title + subject search result is higher than the title search;

3. Row (B/C/D/E,8) gives search result for title + subject after digitizing content and index pages of books under respective class number/s for each respective search terms and is higher than the result of title + subject search before digitizing;

4. Cells (B/C/D,9) give the number of hits for the 229 issues” content pages of 79 Journal titles (Year 2010) search under „Leadership”, „women entrepreneurs” and „disinvestment/divestment” and their synonyms respectively;

5. Cell (E,9) gives the number of hits for the 197 issues” content page search under digital preservation 8 articles and under broader terms 7;

6. Row (B/C/D/E,10 to B/C/D/E,13) gives the relevance ratio for the title; title + subject; the content page; and the index page search of the books respectively and it is observed that it is higher than the only title search result relevance ratio;

7. Row (B/C/D/E,14) gives the relevance ratio of journal content page which indicates that efforts of 1:0.62 is wasted in the manual search as against 1:0.38;

8. Row (B/C/D/E,15) gives the relevance ratio of the result of title + subject for all the search terms after digitizing the content and index pages searched under each terms

<table>
<thead>
<tr>
<th>Row No.</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Journal Content pages (229 of management) and (197 of Library Science) search result</td>
<td>73</td>
<td>1</td>
<td>1</td>
<td>8+7</td>
</tr>
<tr>
<td>10</td>
<td>Relevance ratio for the title search</td>
<td>0.107648725</td>
<td>0.2</td>
<td>0.03030303</td>
<td>0.278481013</td>
</tr>
<tr>
<td>11</td>
<td>Relevance ratio for the title + Subject search</td>
<td>0.175637394</td>
<td>0.2</td>
<td>0.03030303</td>
<td>0.46835443</td>
</tr>
<tr>
<td>12</td>
<td>Relevance ratio for the content page search</td>
<td>*</td>
<td>0.1</td>
<td>0</td>
<td>0.025316456</td>
</tr>
<tr>
<td>13</td>
<td>Relevance ratio for the index page search</td>
<td>*</td>
<td>0.2</td>
<td>0.03030303</td>
<td>0.050632911</td>
</tr>
<tr>
<td>14</td>
<td>Relevance ratio for the Journal content page search</td>
<td>0.3188</td>
<td>0.0044</td>
<td>0.0044</td>
<td>0.07614</td>
</tr>
<tr>
<td>15</td>
<td>Relevance ratio of all the combined search for books made at one go</td>
<td>*</td>
<td>0.5</td>
<td>0.06</td>
<td>0.54</td>
</tr>
</tbody>
</table>

* Study not conducted for those subject/s whose class number search result exceeded 100.
+ For digital preservation the synonyms were not identified, but some broader terms like digital library/libraries were used.
and their synonyms is higher for all the search terms than that of without alternate subject access terms;

9. The search by subject in the OPAC of SOUL 2.0 incorporates Class No. field MARC-21 Tag No. 082$a (for DDC) and uncontrolled terms MARC-21 Tag no.653$a gives an exhaustive list of the documents.

10. The relevance ratio increases with digitized content and index pages and can be projected as 1:1 for the combined search result of the title and subject in SOUL 2.0 OPAC result.

Suggestions

It is suggested from the above observations that giving alternative subject access supports efficient and effective searching. This is achieved by building the query (search string) with the help of thesaurus and acronym master as noted in observation no.1. Ideally user is to be given an option to search primary search term coined by user assuming that the user knows the subject. The search module should be interactive with options like – do you mean this? /would like to search these also? This can be done by giving options as check boxes to select one or more terms from the thesaurus combining with „OR” options in one single field. The Boolean operators can then be used to combine the search in the select field like title, title + subject, author, type of material and year of publication/publisher/place series etc. If the primary term used by the user is not present in the thesaurus it should automatically be added to the thesaurus if the hit list includes this term. There should be a master file of acronyms which while giving search term would include it for searching along with primary search term and all the synonyms.

The Library software should enable uploading of word or excel file (document like text) into the Uncontrolled term field rather than copying from the file and pasting in the field to avoid any error while selecting the text to be copied and deleting already existing text while pasting.

The full record view should highlight the term searched along with the page reference to easily locate on the content and/or index page displayed. The uncontrolled terms field though repeatable, should correlate the search term with volume and part statement of the multi volume document.

The user should be able to save/print the bibliography in the order of his preference (ascending/descending) either by author or title or relevance or by the year of publication.

At the end of search the library software should be able to generate the relevance ratio by computing the ratio of actual documents referred either on-premises or outside-premises to the total number of hits. It should also compute the total number of times document referred for identifying the most popular document for the subject. The user should be given choice of receiving current content services on registration for the service. On receipt of any new document containing the search term information retrieval system should generate bibliography to be mailed to the requester. It should be combined with the feedback of the user to evaluate the services and on receipt of article request update the relevance ratio as well as popularity of the document.

If the publishers provide these details (title page, content pages, index pages and book jackets) for books and journals as word document on CD with each book at the time of purchase or make it available on-line, it will save time; do away with the hardware and technical staff required; overcome all errors that are likely to be encountered with poor quality of scanning and distortion of the text while converting these images into Optically Recognized Characters. However for compiling bibliographies for citations and references development of expert system can be exploited.
Conclusion

It is, thus, seen from the pilot study reported above that the bibliographic database with the key part of the whole document – digitized content, index pages – serving as an alternative subject access gives a wider list of relevant material faster as compared to the search using the only traditional tools. In fact, by this process the library material will become more visible and its use is, therefore, likely to increase. It is suggested that large libraries can consider adopting this model of digitization of the main collection instead of full digitization of select documents or all documents for information search. More user studies on the above line would certainly help in fine tuning this process.

References
Carol L. Tilley - University of Illinois at Urbana-Champaign, USA
Kathryn A. La Barre - University of Illinois at Urbana-Champaign, USA

What if they build it and no one comes? Balancing Full-Text Access and User Tasks

Abstract
Kuhlthau (2005) highlighted the chasm separating the investigation of information-seeking behaviors and the design of information retrieval systems and proposed that its continued existence prohibited any effective resolution of information seeking problems. Consequently she argued that only by enhancing collaboration among researchers engaged in these two areas, continuing investigations beyond the first round of findings, and using findings to design systems for evaluation can research in library and information science proceed meaningfully and effectively. This paper will discuss the task analysis findings from the research project Folktales and Facets that seeks to address this chasm. As part of the content analytical approach taken in this project, the researchers conducted task analysis on transcripts from interviews with 12 subjects who rely on folk narratives in their professional lives as storytellers (n=6) scholars or instructors (n=4), or librarians (n=2). These findings will be mapped to the FRBR user tasks and placed in the greater context of user-focused studies that seek to do the same.

Introduction
Since 2005, Library and Information Science researchers have continually found themselves grappling with a dynamic information environment both in terms of the increasing prevalence of full text resources, and in terms of changes in the codes and standards by which bibliographic access is created. User expectations have shifted accordingly. At the heart of many of these changes lies the Functional Requirements for Bibliographic Records (FRBR) final report (IFLA, 1998). According to this report, the entity relationship model of FRBR intended to:

[P]roduce a conceptual model that would serve as the basis for relating specific attributes and relationships (reflected in the record as discrete data elements) to the various tasks that users perform when consulting bibliographic records. … It takes a user-focused approach in analyzing data requirements insofar as it endeavours to define in a systematic way what it is that the user expects to find information about in a bibliographic record and how that information is used. (IFLA, p.3)

Yet, according to Madison's (2005) description of the origins of FRBR, despite the emphasis on users, the members of the Study Group of the Cataloging Section of IFLA decided not to conduct actual user surveys and instead selected another more expedient and practical option:

[T]o use our collective knowledge of the various types of users from the working group membership and commentators, as well as to draw upon experts in the fields to provide necessary user perspectives and conclusions… Furthermore, the Study Group believed that, given its own international representational make-up, the range of the members' professional cognizance of user needs based on their theoretical and practical backgrounds and experiences, and the types of experts that they expected to draw into their research, would mitigate the need for user surveys (Madison, 2005, p. 29).

The intent of this brief report from the Folktales and Facets project is to provide a glimpse into a real world user survey of information seeking tasks among a small group of scholars, teachers and storytellers as they seek to discover information resources related to their interests. This briefing will compare and contrast observed user tasks with those formulated by the IFLA study group, which relied on the theoretical and practical backgrounds and experiences of the study group members. Because this user study is a small group, no claim is made to the generalizability of the findings; instead, the findings are offered as a response to Barbara Tillett’s (2005) call for more research about user needs
and user tasks in order to provide answers to questions such as, "What information is ‘enough’ for each of the user tasks?" and "What other user tasks are essential?" (p. 198).

**Literature Review**

Today, it is generally understood that the four generic FRBR user tasks—find, identify, select, and obtain—seek to represent a broad set of users, who are motivated by a range of tasks that include searching and/or using bibliographic records across a variety of tools, comprised of "bibliographies, library catalogs, bibliographic databases, full text, or graphical resources" (Madison, 2005, p. 30). Several recent studies attempt to redress the lack of actual user studies in the FRBR final report (1998), including Zhang and Salaba’s (2009) Delphi study with thirty-three field experts, who identified key issues and challenges for FRBR. Among the most pressing issues they identified were the need to verify and validate the FRBR model using real data as well as the critical need for user studies. In another study that included card sorting, conceptual mapping, and comparison tasks, Pisanski, J. & Žumer (2010a, 2010b) examined whether user mental models of the bibliographic universe fit the FRBR entity-relationship model. One important finding of this work is that users do not have shared mental models, and individual mental models often differ from the FRBR conceptual model, thereby making the need for user studies even more evident.

Foster, Clark, Tancheva, & Kilzer (Eds.) (2011) present work that moves user-based observation of the FRBR conceptual model forward immeasurably, at least in the context of creating the eXtensible Catalog. The chapters in this book describe a process of design and development that relied extensively on participatory design and eliciting user responses as a way to identify needed features – including FRBR-based metadata management. Other researchers, including Elaine Svenonius and Barbara Tillett have suggested possible additional user tasks to add to the original four: Find, Identify, Select and Obtain. Svenonius suggests that the Find task be split into two, Locate and Collocate, and added Navigate (Svenonius, 2000, 62-66). Tillett also adds a possible task to the list, Relate (2005, p.198). It is in the context of these offerings that the Folktales and Facets observations of user tasks are offered.

**The Tasks**

The following sections will first discuss the FRBR user tasks for context, and then will introduce the tasks from the project Folktales and Facets.

**The Tasks: Functional Requirements for Bibliographic Records**

The FRBR report (IFLA, 2009) provides a mapping of tasks to attributes and relationships, and generic user tasks. In so doing it also provides a relative indication of the importance of each task to the attributes and relationships of each FRBR entities [Works in Table 6.1 (pp. 88-89), Expressions (pp. 90-92), Manifestations (pp. 93-95) and Items (p. 96)].

**Find**

*Find* is the search for a resource that corresponds to stated criteria (i.e., to search either a single entity or a set of entities using an attribute or relationship of the entity as the search criteria).

**Identify**

*Identify* is to verify the identity or characteristics of a resource (i.e., to confirm that the entity described or located corresponds to the entity sought or to distinguish between two or more entities with similar characteristics).
Select

Select is to choose a resource that is appropriate to the user's needs (i.e., to choose an entity that meets the user's requirements with respect to content, physical format, etc., or to reject an entity as being inappropriate to the user's needs).

Acquire

Acquire is to access a resource either physically or electronically through an online connection to a remote computer and/or acquire a resource through purchase, license, loan, etc.

The following section introduces the tasks from the Folktales and Facets study in order to provide context for the comparison of both sets of tasks to follow.

The Tasks: Folktales and Facets

The user tasks that serve as a comparison to those identified in the FRBR model come from an ongoing study of how individuals seek access to and use folktales and similar materials in their daily lives. With institutional approval, La Barre and Tilley, the principal researchers for this project, Folktales and Facets, have interviewed twelve people. Six of the participants self-identify as scholars (i.e. researchers and instructors); and all but one informant identify as practitioners (i.e. storytellers and librarians). Each of these informants utilizes folktales and similar materials in their work. Full information about this project, nine of the twelve informants, and the method can be found in La Barre and Tilley (2012); however, Table 1 summarizes some of the salient characteristics for each informant.

Through a series of individual hour-long semi-structured interviews, we elicited typical information-seeking and information-use tasks, strategies, and resources related to folktales for these informants. We captured additional information such as the informants' ideas for strengthening discovery and access tools during these interviews, but much of this additional information is beyond the scope of the current paper.

Table 1: Informants Interview for Folktales and Facets

<table>
<thead>
<tr>
<th>Pseudonym</th>
<th>Primary Role</th>
<th>Secondary Role</th>
<th>Highest Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curt</td>
<td>Storyteller</td>
<td>Instructor</td>
<td>M.A., Folklore</td>
</tr>
<tr>
<td>Liz</td>
<td>Scholar</td>
<td>Storyteller</td>
<td>Ph.D., Library Science</td>
</tr>
<tr>
<td>Jeanie</td>
<td>Scholar</td>
<td>Librarian</td>
<td>Ph.D., Library Science</td>
</tr>
<tr>
<td>Alison</td>
<td>Instructor</td>
<td>Storyteller</td>
<td>Ph.D., Library and Information Science</td>
</tr>
<tr>
<td>Glen</td>
<td>Storyteller</td>
<td>None</td>
<td>MLS, Library Science</td>
</tr>
<tr>
<td>Barbara</td>
<td>Librarian</td>
<td>Storyteller</td>
<td>MLS, Library Science</td>
</tr>
<tr>
<td>Rick</td>
<td>Librarian</td>
<td>None</td>
<td>MLS, Library Science</td>
</tr>
<tr>
<td>Sharon</td>
<td>Storyteller</td>
<td>Scholar</td>
<td>M.A., Folklore (in progress)</td>
</tr>
<tr>
<td>Diana</td>
<td>Storyteller</td>
<td>None</td>
<td>M.S., Museum Studies/Education</td>
</tr>
<tr>
<td>Andrea</td>
<td>Scholar</td>
<td>None</td>
<td>Ph.D., Comparative Literature</td>
</tr>
<tr>
<td>Betty</td>
<td>Storyteller</td>
<td>Librarian</td>
<td>MLS, Library Science</td>
</tr>
<tr>
<td>Peggy</td>
<td>Storyteller</td>
<td>Librarian</td>
<td>MLS, Library Science</td>
</tr>
</tbody>
</table>

Four broad categories of information tasks, or goals, emerged from our analysis of the informants' interviews: create, collect, study, and instruct. Each informant engaged in each type of information task. Although the storyteller and librarian practitioners participated in study and instruct tasks less frequently than scholars, the scholar informants engaged equally in all four tasks. The following sections briefly define and illuminate each task.

Create

Create is the assembly and synthesis of multiple pieces of information in order to entertain, elicit an emotional response, or encourage empathy. This activity is most
associated with the storyteller practitioners, such as Sharon who described how she might integrate several versions of a single story to maximize coherence and highlight particularly memorable language or humor. Betty and Peggy, who sometimes perform stories as a duo, discussed how collaboration can be a creative activity, as it requires agreement on source materials, timing, and performance aesthetics.

Collect
Collect (or acquire) is defined as gathering and organizing information resources or their surrogates for current and/or future use. The storyteller practitioners provided some of the best examples of the collect task. For instance, Curt has created a personal library of more than 3,000 volumes, which he has cataloged using Library. Thing and makes available for students and storytellers. In addition, this library, together with extensive performance notes, story outlines, and similar resources that Curt gathers electronically on this computer, serve as the starting point for many of his scholarly projects. Other informants collect items such as Internet bookmarks, set lists, and audiotapes; to gather and organize their collections, the informants use a variety of tools including paper folders, map cabinets, file drawers, guitar cases, and notebooks.

Study
Study is the acquisition of information in order to transform intellectual understanding. It differs from create in that the transformation and synthesis of multiple pieces of information is targeted primarily at creating new knowledge as opposed to eliciting an emotional response. For many of our informants, folktales themselves are not the objects of study. Jeanie, for instance, has a long-term scholarly interest in the pioneering storytellers in the early decades of youth services librarianship. Scholar Andrea focuses her work more on the material culture surrounding the production of fairy tales. Similarly, Liz recently completed an edited book in which women in academe reflected on the stories that have motivated their professional work.

Instruct
Instruct is the assembly and presentation of information to train others in some skills, knowledge, or behavior. As an example, Alison regularly instructs students in graduate-level storytelling classes in the use of particular tools and strategies to identify folktales for re-telling. For her job as a museum educator, Diana frequently develops story performances designed to inform young audiences about the social and cultural contexts of various museum exhibits.

Strategies Related to Tasks
The informants relied on two overarching strategies to works toward their goals: search and browse. As many of the informants had formal training and work experiences as librarians, their search strategies often belied this background. For instance, Betty discussed limiting a search in a library OPAC by publication date for a quasi-known item (i.e. an object that the informant knows exists even though traditional known-item search information such as title and author are unknown). Pearl-growing and complex keyword-chaining search strategies were common among participants as well. For all informants, browsing was the primary information-seeking strategy used to fulfill the four tasks. Many of the informants, for example, relied on browsing in a library by call number or on the displays of new items to identify promising stories or collections.

Tasks: Compare and Contrast
The FRBR user task Find most closely resembles the project participants' Search strategy in support of a variety of tasks, or goals. Here, both Search (Folktales and Facets) and Find
(FRBR) involve a search for relevant materials that fit a user's stated search criteria. For the participant group, search criteria may be a theme or a particular source, it is quite rare for these users to look for a specific author, or title, though far more common for them to look for items on a given subject or theme.

The *Folktale and Facets* participants also relied on an *Explore or Browse* strategy to help locate materials - especially in cases when they seek variants or adaptations of a tale. Here the *Browse* strategy more closely resembles Svenonius' suggested FRBR task *Navigate*, than *Find* (or *Search*). When resorting to the *Browse* strategy, a participant may not be able to articulate exact search criteria, but is easily able to recognize an item or 'know it when they see it.' Both *Browse* and *Navigate* bear a close resemblance to Tillett's suggested FRBR task *Relate*. The participants' *Browse* strategy is heavily reliant on associative relationships in order to accomplish tasks. These associations may involve related sources, items in aggregations, classification number associations or other pathfinder-like devices such as folktale indexes that participants often use in order to locate adaptations and variants of tales in ways that are not well supported by the metadata in commonly provided in bibliographic records.

The *Instruct* and *Study* tasks of the *Folktale and Facets* participants seem to be most closely related to the FRBR task *Select*. When participants engage in the *Instruct* task, they are preparing materials or programs that will teach others. When participants engage in the *Study* task, they are usually conducting original research or editing resources that are related to folktales, storytelling or folk narrative. Both tasks rely on materials that are identified through the FRBR task *Select*, but move beyond mere selection to the act of analysis and/or synthesis of the selected information resources. Choosing the appropriate resources (*Select*) is just the beginning phase of these two tasks, and it is the end goal that distinguishes between the two. The *Instruct* task often relies on the *Browse* strategy, while the *Study* task more often relies on *Search*.

*Folktale and Facet* participants also engage in the *Create* task, in which they prepare songs, notes, pathfinders, or programs that are primarily intended as entertainment, rather than as instructional devices. Many participants talk about the importance of being able to find numerous variations or adaptations of a story or variations of material that fit a particular theme as they work towards the creation of a version that they will make their own. Here the FRBR task *Obtain* acquires a central position as the participants must not simply identify the necessary items, but must also be able to acquire them.

This leads into the last *Folktale and Facets* user task, *Collect*, which also seems to extend beyond the scope of the likely companion FRBR task, *Obtain*. The *Collect* task includes more than the act of acquiring or obtaining resources in that it is primarily driven by the desire of participants to actually build physical or virtual collections of materials that they will draw from over time in an open-ended fashion. Nearly all of the participants of this project speak of their personal collections of materials, some of which they have created and others which they have purchased. They also speak of a desire to share their collections with others and are eager to find ways to do so.

This mapping shows a high degree of interplay between and among the tasks and strategies of the *Folktales and Facets* users. It also suggests several refinements and extensions to the four basic FRBR tasks, while providing support for *Navigate* and *Relate* as additional tasks. What this mapping also indicates is that the tasks for specific user groups may be far more nuanced than is possible within the four main FRBR tasks, especially without the addition of the *Relate* or *Navigate* tasks. The FRBR report (1998) does not indicate that the number of user tasks should remain at four, and it remains open to the need for a variety of tasks in the model, among them preserve, and management of rights. The
preliminary findings from this project indicate support for additional, and more finely nuanced descriptions of user tasks, as well as the need for further attention to Tillett’s admonition ‘What information is enough.’

One possible explanation for the nuances discussed in this briefing are addressed by Nicholas (2005) as a set of considerations for oral narratives, a format currently not well treated by bibliographic access forms. This weakness stems from the understanding within the oral tradition that an audience experiences a new version of a tale on each telling and that a tale relies not only upon the source material used by the teller, but also upon audience members’ common memory and common knowledge of previous versions of the same work. The teller depends on this common shared memory as he introduces variations, which may concern the elements including plot, characters, and title. (p. 180). Bibliographic access forms typically deem such variations too superficial to capture, even though they lie at the heart of the oral narrative experience. Another feature of oral tradition that must be considered when contemplating the question of ‘What information is enough’ is the phenomenon of ‘sunken works.’ These often consist of written recollections of an oral tale in a colonialist work. These composite works result in ‘bibliographically dependent or hosted works’, which sink out of sight in bibliographic records, especially when the tale belongs to a group of people who are now extinct (pp. 189-190). Nicolas looks to the FRBR model as a way to redeem these ‘sunken works.” In combination with the user-focused tasks that are indicated by the Folktales and Facets project, further support for the importance of user studies is evidenced.

Conclusion

Today, those who seek to provide access to information resources are more and more frequently resorting to heavy investment in full text portals. Individuals, like those in the Folktales and Facets project are being directed to the purported wonders of portals like the Open Folklore Initiative (http://openfolklore.org/). This site provides access to tools that allow a user to search across HathiTrust-scanned materials, which include the full collection of Indiana University’s unparalleled folklore materials. Unfortunately not everyone can actually view the full text of these materials due to access limitations; at a minimum, they can identify if and where terms appear in an item and see a small snippet of context for those terms. Studies like Folktales and Facets present a preliminary set of user tasks that could come to bear on the development of user-focused features for full text access portals that can also benefit from enhanced understandings of the FRBR entities, attributes and relationships between and among folk narrative materials.

References


Users’ Perception of Aboutness and Ofness in Images: An Approach to Subject Indexing Based on Ervin Panofsky’s Theory and Users’ View

Abstract
It is widely accepted that subject indexing of an image is based on a two-dimensional approach. The first is the ofness and the second focuses on aboutness of the image. Assigning a suitable set of subject tags based on these two groups depends, to a great deal, on users’ perception of the image. This study aims at analyzing users’ perception of aboutness and ofness of images. 25 in-depth semi-structured interviews were conducted in two phases. In the first phase a collection of 10 widely known photographs were given to the interviewees and they were asked to assign subject tags (as many as they wanted) to each image. In the second phase some facts regarding each image were given to him / her to assign further tags (again as many as they wanted) or even modify their previous tags. The results show that the interviewees do focus both on ofness and aboutness in subject tagging; but it seems that they emphasize more on aboutness in describing images. On the other hand, as soon as the interviewees were able to distinguish the iconographical ofness, they could speak of iconographical and iconological aboutness. The results also show that subject indexers must focus on the iconographical level, especially regarding those tags which represent the ofness at this level.

Introduction
Subject indexing of images is a matter of problematic nature from different perspectives. The first and may be the most important issue, deals with the lack of standard or universally-applicable code for a meaningful subject representation of an image. Although it is widely accepted that subject indexing of an image is based on a two-dimensional approach (i.e. ofness and aboutness description), the problem still lingers in determining the depth of descriptions especially when it comes to aboutness. Based on ofness/aboutness approach the subject tags to be assigned to an image could be divided into two different groups. Those which describe the things or objects depicted in the setting of an image constitute the ofness; and those tags which describe the matter of theme or the story expressed in an image fall into the aboutness group. Most of the literature in the field of subject access to images focuses mainly on the works of art and other cultural heritage objects. Yet there are issues regarding the whole process of subject description of images (moving or still) (Yee,1990; Shatford lane, 1994, 2002; Kessel, 2011). Yet other pictorial works with special characteristics are being neglected. News images or scientific figures are among them. This paper attempts to address the problem from a different perspective (users’ point of view), based on a different test bed (News Images). Thus this study aims at analyzing users’ perception of aboutness and ofness of news images. Users’ perception is of significance because it shapes the keywords through which users decide to search any image database or even the Web. In order to set a framework for analyzing users’ perception, ofness / aboutness of the images used was interpreted in the light of Panofsky’s theory.

Ofness/ Aboutness and Panofsky’s Theory in Iconography
As Shatford Layne (2002) puts it, subject access to images is among the most important means through which users can find their desired pictorial works. From this point of view there are two main aspects in subject interpretation of pictorial works; ofness which considers what is depicted in an image and aboutness which deals with the interpretation of the subject matter and identifying what has been symbolized in an image. As Yee (1990)
mentions, this aspect covers a broad territory due to the fact that it includes expressional and latent meaning. Thus index terms for images are assigned based on these two aspects.

Although different scientists adhere to the fact that differentiating ofness and aboutness in images is an accountable way for interpreting the subject and these two dimensions do give us the clues to select the correct terms for subject tagging of pictorial works, there are other specialists such as Svenonius (1994) who believe that it is almost impossible to express what is symbolized in one medium in another. In other words, it is difficult to adequately index visual and pictorial works by means of verbal language. The dilemma is only solved (at least to some extent) if both subject indexing knowledge and a bit of visual/verbal symbolism knowledge are mixed together. Here rises another point regarding the subject matter in works of art based on Panofsky’s theory in iconography, since iconography as a branch of art history deals with the subject matter or meaning of works of art. Panofsky (1962) identifies 3 different strata for subject meaning of pictorial works: 1) The Primary subject matter which is identified through pre-iconographical description. In order to develop such a description practical experience (i.e. familiarity with objects and events) is required. Identifying lines, colors, basic shapes, objects and etc. takes place at this level (Peters & Stock, 2006). 2) The secondary or iconographic subject matter which is identified through iconographical analysis. This type of analysis requires knowledge of literary sources (i.e. familiarity with specific themes and concepts). Events, Personas, specific objects and concepts are to be identified at this level. It is here that the interpreter is able to explain what is really going on in an image. If the image’s icons are correctly identified then its true story is revealed. 3) The intrinsic meaning of the image (Iconology) which is identified through iconographical interpretation in a deeper sense (i.e. Iconographical synthesis). Developing such an interpretation requires synthetic intuition which is familiarity with essential tendencies of human mind and it is conditioned by personal psychology and “Weltanschauung”¹. At this level things which have been inferred from the image are not necessarily depicted in it.

Arriving at the intrinsic meaning of an image requires correct pre-iconographical description and then correct iconographical analysis. The first phase is simple. It deals with lines, colors, basic shapes and as long as we have the practical experience of encountering the motifs, we would certainly be able to distinguish them. Yet Iconographical analysis phase focuses on the stories and themes (Sieger, 2010). It requires the capability to derive right conclusions regarding the aboutness of an image in the primary level; the stage when one can name the icons and say what is really happening in the work of art. Yet interpretation of the intrinsic meaning of an image is based mainly on the symbolic interpretation, and requires the knowledge and ability to make sense of what has been expressed behind those depicted. This skill is equal to the capability in inferring the aboutness of the image in a more elaborate level. Fujita (2000) in a paper describes this aboutness at primary level and elaborates level as the foreground image where main motifs are placed and background images where clues are hidden in order to give implicit information in the picture.

Methodology

For the purpose of this study a qualitative approach has been used. 25 in-depth semi-structured interviews (each session took at least 30 minutes) were conducted in two phases. In qualitative research the sample size is determined by data saturation which means new categories, themes or explanations stop emerging (Marshall, 1996). As for this study, after the 20th session there was no sign of new keywords or even interpretations (right or wrong) from the interviewees. Thus the interview sessions were stopped with the 25th interviewee.

¹ World-View
In the first phase a collection of 10 widely-known news images were given to each interviewee and they were asked to assign subject tags (as many as they wanted) to each. The images depicted 10 famous events or situations in Iran (5 pictures) and the world’s history (5 pictures). If the user by chance was not familiar with the theme or the true story of the image, some facts regarding each image were given to him/her for further tagging (again as many as they wished) or even modify their previous tags. Since these images were selected from award-winning or well-known news images, and since the audiences of news images were ordinary people, the interview sessions were settled in a park where the researcher could find ordinary people with different backgrounds. Azadi Park, the largest park (area: 200 km$^2$) in Shiraz city (Iran) was selected as the setting for the data gathering sessions. The interviews were recorded and the tags were also written down on the two part form designed for writing down the tags and descriptions made by the interviewees during the first and second phases of the interview.

Analyzing the interviewees’ responses, the tags were divided into those which represented the ofness and those which represented the aboutness concepts and then basic features of Panofsky’s Theory of Three Strata of Subject Matter which have been previously enumerated in a short checklist were mapped on both groups of interviewees’ tags (those reflecting ofness and those reflecting aboutness).

**Findings**

**Aboutness Tags vs. Ofness Tags**

The results show that the interviewees focus both on ofness and aboutness of an image in subject tagging; but it seems that about two-third of their subject tags are assumed to reflect the aboutness of each image. For instance regarding Fig. 1, for every single tag focusing on ofness, 2.2 tags regarding aboutness of the image was assigned. Popular ofness tags assigned to this image were “Two Hands”, “Black hand”, “Africans”, and “Ugandan Child”. Popular wrong ofness tags were “Somalies”, “Somali child”, “Somalia”. Although these tags share some semantics and symbolic icons with the main idea of the image, wrong tags here best denote the important effect of familiarity with themes based on the literary (i.e. recorded) knowledge in Panofsky’s theory. Recently Iranians have been bombarded by myriads of news about the drastic situation of Somali people. Thus first impression of some of them was that this picture is actually depicting a Somali child. Next, when the real date and theme of the image was revealed to them, they just kept correcting their previous tags regarding the time and also the place. In other words, after explaining the theme or story of each image, most of the interviewees began to correct their iconographical impression. Yet their tags of aboutness remained mostly unchanged.

**Fig. 1: Uganda, 1980, A Starving Boy and a missionary**

A deeper speculation of the interviewees remarks on images, reveal this interesting point that although inferring aboutness at the level of iconological interpretation is almost impossible without correct iconographical inference of works of art, in case of news images it is possible to infer just the right aboutness even at the highest level in Panofskys’ three strata without knowing the true story of the image (i.e. when the icons in the image are not
known). In other words based on table 1, it is possible to have pre-iconographical tags and then jump to iconological tags. One of the interviewees regarding Fig.1 mentioned five tags “Thin, wrinkled black hand” and “white hand”, “Famine”, “Capitalism” and “Need for international help”.

To assign these last two tags to the image is a matter of the indexing database policy, but from the users point of view they are not wrong; and there might be somebody in search of a news image similar to this for fund raising campaigns.

**The Effect of Demographic Variables in Tagging**

During the data gathering sessions the researcher tried to balance the distribution of demographic characteristics of the interviewees. 13 female and 12 male interviewees took part in the study. Their ages ranged between 16 and 76. Based on the interviews, no evidence regarding any particular difference between male and female participants was found. But as for the age, some delicate differences were perceived. Those who were 30 or older acted better at iconographical level, both for tagging the ofness and especially aboutness. Scrutinizing their characteristics revealed more. Those who were more disposed to different types of mass-media were better in iconographical interpretation, yet those who were younger and used internet frequently were more familiar with the concept of tagging and thus they were more selective. It seems iconographical interpretation in news images requires more general information. In case of Fig. 2, there were a few participants who failed to recognize Mosaddeq. In such cases their guess regarding ofness was “Thinking man”, partially correct guess on aboutness was “thinking”, wrong impression on aboutness at the iconological level was “Age” and the worst tag on ofness was “Monk”.

![Fig. 2: Mohammad Mosaddeq 62nd Prime-minister of Iran](image)

General information plays a vital role here since different interviewees with age from 19 to 76, who recognized the icon depicted, were in different educational levels, and yet the one who made the worst choices was a Chemistry student. Thus one can infer that it is wrong to have total confidence in users tagging, unless a clue is provided, maybe a caption or a short summary for the image; and there must be some sort of regular checking.

**Gradual Interpretation**

Based on analyzing the tags provided for each image by interviewees, a matrix was prepared showing the confluence of aboutness and ofness tags as columns and Panofsky’s three strata for each image. Based on the interviews the discerning sequence, through which each group of keywords is provided, could be developed. Table 1 depicts a merged matrix with the discerning sequence for each group of keywords.

---

2 Here “famine” is regarded as a tag regarding iconographical aboutness of image.
3 The best preliminary interpretation on ofness was given by a small kid (~6) who was playing in the park while I interviewed her mother. She came, saw the picture and said: “this is a small black hand with red sleeve in a white hand”. None of the interviewees mentioned the depictions in such details, but anyhow she was not assumed as a participant.
Table 1: Mapping matrix for Panofsky’s theory and Ofness/Aboutness description

<table>
<thead>
<tr>
<th></th>
<th>Ofness</th>
<th>Aboutness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iconography</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iconology</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For a better understanding, the results for Fig. 3 are discussed here. This image depicts a demonstrator confronting a line of People’s Liberation Army tanks in China (1989 - Tiananmen Square)\(^5\).

Fig. 3: 1989- Tiananmen Square Demonstrations

During the interviews it was observed that at the first sight there is a tendency to tag based on objective impression. (For instance: “Tanks”, “People”, “street”, “A man in front of tanks”, “Army”, “Tanks in the city”). At this stage only the objects depicted directly in the image are mentioned. Gradually those who were not familiar with the theme of the image, expanded their description to tags such as “Military Parade”, “war”, “Civil war”, “Invasion to country”, “demonstration” at first and then to more symbolic terms such as “Resistance”, “Sacrifice”, “Life vs. death”, “Faith”, “Freedom”, “Hegemony”, “Forced to fight”, and “capitalism\(^6\)”. As it is evident in these three groups of tags, two types of aboutness are to be inferred. The tags regarding ofness at this stage are well-classified under preliminary interpretation of ofness. Those aboutness tags inferred from the objects including terms such as military, war, or demonstrations could be classified under preliminary/ pre-iconographical interpretation of aboutness. But the latter group of aboutness tags is symbolic enough to be classified under iconological interpretation of aboutness. Wrong interpretation of aboutness at this level is inevitable as it is evident in both groups of aboutness tags.

The next stage is the iconographical interpretation of aboutness. Some reached this level (4 persons) but others asked the interviewer to tell them the image’s theme. Among those four who guessed the theme 2 were right. One of them exactly mentioned Tiananmen and the other person mentioned China, and both of them enhanced their tags with the date in the second phase of the interview. Those who gave wrong tags mentioned “American tanks” and “Israeli tanks”. At this level the correct icons are identified. And only based on a correct iconography, a correct aboutness in iconographical and iconological levels is

\(^5\) 1\(^{st}\) prize world press photo, by Charlie Cole.

\(^6\) This is due to wrong preliminary interpretation of ofness. The interviewee failed to see the red stars on Tanks. It is important to note that the interviewees were ordinary people who might not have proper reading rate or enough general information.
reached. At this level ofness tags such as “Chinese demonstrator”, “Tank-man”, and “Chinese army tanks” are provided. Iconographical aboutness tags such as “Tiananmen demonstrations”, “Tiananmen square demonstration”, and “1989 China” were provided afterwards and totally wrong tags in iconological aboutness level mentioned prior to the correct iconological notion of the image such as “capitalism” changes to “communism”. In table 1 the whole sequence of interpretation is shown through numbered arrows. To sum up, we infer that one can interpret objects in a picture and infer the aboutness of it. At this stage if the observer develops a false impression from what has been depicted, false aboutness is inevitable. According to Panofsky it is hardly possible for a person to base a correct iconological impression on a pre-iconographical inference. This study shows that at least regarding the news images this doesn’t always hold true, since from the primary stage one can infer to some extent the iconological aboutness of the image (Table 1). On the other hand, as soon as the interviewees were able to distinguish the iconographical ofness of the images, they could speak about iconographical and iconological aboutness.

Which Group of Tags to Prefer?

Another important question which remains is the significance of each group of keywords for better retrieval. Each group of tags was searched in “Google Image” and the results showed that by utilizing those tags which deal with ofness and aboutness at the Iconographical level of description, the desired image is returned within the first page of the retrieval. Further investigations revealed that those tags regarding pre-iconographical interpretation of ofness and aboutness are weak in retrieving the image within the first page of retrieval (for 50% of images the results appeared in the 3rd or up). As for those symbolic tags the search results were hopeless. But the interviews showed that the number of people who exactly know the theme of the picture and thus those who could correctly interpret the iconography of image elements is not large. Also due to the fact that people are not eager to give tags regarding ofness, it would be a reasonable decision to assign some elemental primary-level tags regarding the ofness to each image. Based on the tags that interviewees assigned, it is evident that aboutness tags at iconological level could be sufficiently provided by users; but since personal judgment highly affects the decision regarding the keywords at this stage and because these tags are not successful in providing precision, it is better to keep these groups of tags as small as possible. This means that a subject indexer is needed to assign/add necessary tags regarding iconographical ofness and aboutness and keep the amount of tags regarding iconological aboutness small, since users are weak in this regard. During this process, special favor should be given to those tags which are classified under iconographical ofness.

Concluding Remarks

Subject indexing of images is important as people show more interest in image objects on the Internet and in databases. An important point is that interpretation of news images is to some extent different from the works of art. Whether this difference is because of their simpler or less symbolic nature or due to the subjects depicted in them, different audiences with different backgrounds do understand (at least to some extent) them, and their meanings could be conveyed. The whole process of absorbing the deepest symbolic meaning of each image is gradual. During different phases of this process different notions of ofness and aboutness are perceived. On the other hand, general information/knowledge plays an important role during interpretation of the meaning of an image. More general knowledge leads to better iconographical interpretation. This means that knowledgeable people, who have access to information media/sources, assign better iconographical tags regarding ofness and aboutness; and this helps them in developing more precise idea of iconological aboutness. Also it seems that iconographical tags on both ofness and aboutness retrieve better results. But based on the findings of this study, users are not always familiar
with icons in an image. The role of subject indexer or in some systems the image provider (the one who uploads the image) would be of great help through providing iconographical ofness/aboutness tags, or providing a short summary and/or a caption for the image.

This paper opens an insight into assigning subject tags for images based on two well-known approaches of Panofsky’s theory and Aboutness/Ofness description of images. Further research can be carried out to find out the process of interpretation of images by people and the factors that may affect this process. Also we need to find out about those tags assigned by people that are not relevant to the images but represent subjects they think suitable.

References
Communities of Practice, Gender and Social Tagging

Abstract
Social or collaborative tagging enables users to organize and label resources on the web. Libraries and other information environments hope that tagging can complement professional subject access with user-created terms. But who are the taggers, and does their language represent that of the user population? Some language theorists believe that inherent variables, such as gender or race, can be responsible for language use, whereas other researchers endorse more multiply-influenced practice-based approaches, where interactions with others affect language use more than a single variable. To explore whether linguistic variation in tagging is influenced more by gender or context, in this exploratory study, I will analyze the content and quantity of tags used on LibraryThing. This study seeks to dismantle stereotypical views of women’s language use and to suggest a community of practice-based approach to analyzing social tags.

Introduction: Tagging Language and Gender
Social or collaborative tagging enables users to organize and label resources on the web. Users attach textual labels, or tags, to online resources such as photos, articles, videos or websites. Users tag to create personal organization systems, record and share content of online resources, or make recommendations. Information organization researchers have questioned the ability of professionally-assigned subject access terms to reflect the language of users and have proposed that social tagging could bridge that gap to enhance retrieval (Lu, Park and Hu 2010, Adkins, Bossaller and Thompson 2009; Rolla 2009). Furthermore, the dominant controlled vocabularies and classification systems have been criticized as being inherently biased. Libraries and other information environments hope that tagging can complement professional subject access with user-created terms. But who are the users who tag, and are they representative of the user population? Virtually no research has investigated the identities of taggers and the differences in how they tag, mainly because websites do not always require users to provide or make public personal information such as gender, ethnicity, or even real names, thereby making it difficult for researchers to determine demographic information. But does demographic information actually influence how linguistic variation in tags? In terms of gender, research has been split on this topic, both asserting and refuting the existence of a “woman’s language.” Since social tagging is purported to be the voice of the user community, this exploratory study will investigate whether the linguistic characteristics of social taggers divide along gender lines on the website LibraryThing, an online, “full-powered cataloging application,” comparable to a library catalog (“About LibraryThing” 2011). An alternative framework for analysis will be proposed, using the community of practice approach, first articulated by Lave and Wenger (1991) and Wenger (1998) and further developed in the context of language and gender by Eckert and McConnell-Ginet (1992, 1999).

A number of assumptions have been made about the differences between how men and women use language, much stemming from Robin Lakoff’s (1973, 49-51, 1975) work raising awareness of gender differences in language. For example, she believes women use more tag questions, hedges, polite grammar, and intonations that betray insecurity and subordination. She believes that language reflects a power differential, with women working from a self-perpetuating deficit that is unrewarded by society. Though Lakoff’s work was not informed by rigorous research and has since been disputed, her work continues to be uncritically cited. Sunderland (2006, 95-96, emphasis in original) writes that the “use of her findings now entails a theoretically dated essentialist representation of gender; women talk like this, men like that” without any regard of questions of power,
individual nuancing, social and linguistic development and change.” Nonetheless, many of her conclusions have been accepted by the popular mainstream as well as some linguistic scholars, and have formed a core of stereotypes about women and language use.

**Communities of Practice**

More recently, sociolinguists have begun to see gender as one ingredient that affects language use rather than as a constitutive force. Eckert & McConnell-Ginet (1992) criticize generalizations drawn from statistical differences that interpret gender as the reason behind linguistic variation, and instead propose that “communities of practice” influence how people speak. Communities of practice were first introduced in the context of apprenticeship education by Lave and Wenger (1991), and Eckert and McConnell-Ginet (1992, 464) expand its boundaries to include social groups with linguistic and behavioral similarities, or “An aggregate of people who come together around mutual engagement in an endeavor. Ways of doing things, ways of talking, beliefs, values, power relations—in short practices emerge in the course of this mutual endeavor.” Membership is determined by goal and context, rather than inclusive of a demographic group that has no “mutual endeavor.” Communities of practice are both self-reinforcing and transformative in that participants affect the larger community just as the community affects the participants’ practices (Wenger 1998, 59). A person can interact with multiple communities of practice in a single day. Membership may be institutionally reinforced or self-constituting, and gatekeeping can be policed by formal or informal rules of legitimacy. Wenger (1998, 130-131) also includes other characteristics that includes “a shared discourse that reflects a certain perspective on the world,” “shared ways of engaging in doing things together,” and “jargon and shortcuts to communication,” among others.

Dubé, Bourhis and Jacob (2006) extended the concept of a community of practice to the virtual environment, creating a typology of organizational online communities, which includes the broad categories of demographics (meaning of the site, not of its users), context (the environment in which it was created/exists), membership characteristics, and reliance on technology. The original typology was then expanded by Hara, Shachaf and Stoeger (2009) to more specifically define the parameters of each as well as to include open-access communities, or communities that are not institutionally constituted.

**LibraryThing as Community of Practice**

Is LibraryThing a community of practice? Wenger’s (1998, 73) original three requirements—mutual engagement, a joint goal, and a shared resource pool—all are present in LibraryThing, although to what degree could be disputed. Mutual engagement means that members of a community of practice have regular interaction. Through the most direct social aspects of LibraryThing, users can access other users’ libraries, track other users, be tracked, rate books, make recommendations, and join discussion groups. However, though users can participate without directly interacting, those who tag, even if they are doing it for themselves, end up passively interacting every time they view a tag cloud, read a review, or post a book to their library. While there may not be one overarching institutional goal for the entire population of LibraryThing, all users whether fully or peripherally participating, to some degree are invested in the activities offered by the site, whether it means uploading a particular part of their collection, or sharing with others; otherwise, they would not access it.

The community members must share a resource pool, which can lead to questions of power. Davies (2005) discusses concerns of power within communities of practice, arguing that despite a shared enterprise, peripheral or marginal members may be denied legitimacy by more powerful group members, although Lave and Wenger (1991, 36) deny the
existence of a “single core or center” of power. Power differentials can exist in several capacities on LibraryThing. Users join free and have access to hundreds of thousands of book profiles from Amazon and the Library of Congress. Some users pay a one-time twenty-five dollar fee to upload an unlimited amount of books, whereas the first 200 books are free for everyone, providing some users more opportunity to influence the practices of the group. Additionally, people can be granted titles such as “top tagger” or “author member,” which might be additional prestige markers that add to the tagger’s authority. However, the tags are aggregated without identification unless the tags are searched via the personal libraries of those who have tagged it. Since LibraryThing is not a face-to-face community, the concept of “participation” would have to be defined because tagging does not have the ephemeral quality of speech. A user could tag once and never return to the site, but still be considered a member because the words remain. On the other hand, few methods of gatekeeping exist that could reduce legitimacy (for example, reviews can be flagged, but tags cannot and are not censored).

Methodology
Although spoken conversation and social tagging are not direct analogs, since interruption, turn-taking, grammar and the oral qualities of language are not present in tagging, some interpretation of speech behaviors in language and gender research can be applied to social tagging. A number of hypotheses could be formulated about the way women and men should tag, based on “folklinguistic” tales about gender and language are true, meaning the ideas spread by popular culture and media about how men and women speak (i.e. men talk sports, women nag.) (Speer 2005, 30). Applying a selection of these folklinguistic beliefs to tagging behavior, it could be conjectured that women may use more affective tags reflecting women’s ostensibly emotional nature, their need for “intimacy” (Tannen 1990, 26) or use “empty adjectives” marked as feminine, such as “divine” or those that do not convey strong emotion (Lakoff 1975, 51). Women may use more relational/prestige tags to demonstrate their connections to others or a lack of confidence in their opinions. Women might use more specific language (Lakoff 1973, 49) or perhaps more tags to show that women are more talkative than men (Speer 2005, 30). Women may use a totally different way of tagging that reflects a “genderlect” (Tannen 1990, 42). Women may practice more correct spelling or capitalization in the public forum of tagging, as equivalent of using more formal grammar or pronunciation to make up for powerlessness (Coates 2004, 52).

Gender
The term “gender” has a disputed meaning in feminist theory, ranging from essentialist biological sets to constrained performance to linguistic activity to psychosocial traits. For the purposes of this study, gender is defined as a performance of a particular gendered identity, not necessarily related to biological sex. The study relied on the self-disclosed gendered traits to differentiate men from women. The data collection depended on an electronic representation of feminine or masculine identities on user profile pages, without the ability to discern the users’ biological status. As in any computer-mediated environment, the assumption is that users are who they say they are, to the extent that one reveals online. Therefore, the terms “men” and “women” will be used to indicate gender rather than sex.

Data Collection
The tag sample consisted of 414 tags corresponding to eight popular fiction books. They were created by sixteen taggers, split evenly between men and women, who tagged all or
most of the selected books. The books were selected for their popularity to ensure they had been tagged by a number of people of both genders. To identify the gender of the taggers, the user profile pages were consulted. As discussed above, all members have a public profile where information about the user and their personal library can be found. Some users fill out the profile in great detail, describing reading habits, life philosophy, and personal interests. Often users upload photos or other graphics they believe represent them. The profile does not have a field for indicating gender; however, the username, photos, or contextual clues, such as the "real name" field, can indicate one gender or the other. Taggers were not included if any question remained about what gender they were trying to portray.

The selected books were Dan Brown’s The Da Vinci Code, Stephen King’s Christine, F. Scott Fitzgerald’s The Great Gatsby, Seth Grahme-Smith’s Pride and Prejudice and Zombies: The Regency Romance, Now with Ultraviolent Zombie Mayhem!, J.K. Rowling’s Harry Potter and the Sorcerer’s Stone, Alice Sebold’s The Lovely Bones, Gail Carson Levine’s Ella Enchanted, and Audrey Niffenegger’s The Time Traveler’s Wife. The first four could be considered "unmarked," or appealing to either gender, and the second four could be considered "marked" or appealing to the subset of women, even though they were read and tagged by men (Eckert & McConnell-Ginet 2003, 21). Users upload the profiles of the books and can choose assigned subject headings or create their own tags. Therefore, only fiction was used because typically libraries do not assign subject headings to fiction. LC assigns headings for some popular books, but their subject headings are easily identifiable because of their complexity and structure.

Limitations
Several limitations exist. The sample sizes of books and participants are too small to generalize behavior. Similarly, the users made their identities known either by uploading photos or by self-disclosing. Those who are anonymous may tag differently. Also, based on the information publicly available, it is difficult to discern any other identifying social markers. Therefore, the community of practice is the only common denominator for the selected taggers other than gender. Finally, it is unclear whether social tagging is comparable to the type of linguistic activity in communities of practice as discussed by language and gender scholars.

Findings
Quantity of tags
As shown in Table 1, the total number of tags for the eight books was similar: 214 for women and 199 for men. Individual taggers tended to tag at consistent levels; that is, people who supplied only one or two tags for a book tended to do so for all the books. In this set, the men (N=8, M=4.9 SD=3.8) and women (N=8, M=4.7 SD=3) both tended to the extreme ends with women only slightly cleaving closer to the mean. Median is 4 for men and 4.7 for women. Harry Potter had a higher average amount of tags, since taggers tended to list the many fantasy aspects, such as wizards, spells, and magic wands and because it covered several genres (fiction, young adult, fantasy, adventure, etc.). Pride and Prejudice and Zombies averaged fewer tags, probably because the title neatly sums up the book’s subject. Ella Enchanted earned an unusually high average of tags from the men (7.3), but the results were skewed because the three men who tagged the book were characteristically enthusiastic taggers, which brought the average up. Ultimately, no difference was evident regarding how many tags men and women assign to the books, and the different levels of tagging were determined by the individual rather than by gender.
Table 1: Tag Statistics

<table>
<thead>
<tr>
<th>Tag type</th>
<th>Percentage/Raw</th>
<th>Broad/Narrow</th>
<th>Percentage/Raw</th>
<th>Broad/Narrow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genre</td>
<td>38/82</td>
<td>63/37%</td>
<td>30/60</td>
<td>52/48%</td>
</tr>
<tr>
<td>Subject</td>
<td>31/66</td>
<td>50/50%</td>
<td>21/41</td>
<td>36/64%</td>
</tr>
<tr>
<td>Descriptive</td>
<td>10/21</td>
<td>22/43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>6/14</td>
<td>7/16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task</td>
<td>7/16</td>
<td>6/13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>5/10</td>
<td>5/11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relational/Prestige</td>
<td>1/2</td>
<td>6/12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affective</td>
<td>1.5/3</td>
<td>.5/1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Undetermined</td>
<td>.5/1</td>
<td>.5/1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Types of tags**

Much research has focused on what semantic types of tags are most commonly used. Lawson (2009, 577) divided tags into objective or content-oriented tags, and subjective or user-oriented tags. Kipp (2009) identified three of the most commonly found tags on three major bookmarking sites. Subject tags supply information about the contents of the websites or articles (“dogs”). Affective or emotional tags record how the resource makes the tagger feel, and can also serve an evaluative function (“funny”). She also identified time, task or project-related tags, which tend to be for personal record-keeping (“toread”). Other commonly used tag types found on LibraryThing related to book discourse include “genre,” which describes the subcategory of fiction of each book (“suspense”); “descriptive” discusses details about the item unrelated to content (“made into movie”); “relational/prestige” tags relate the item to what other people thought of the item (“recommended”); and finally, some tags were so personal that no meaning could be discerned from them (“4cot”). These were noted as “undetermined” in Table 1.

Table 1 shows percentages of tag types used by men and women. Both genders followed a similar Zipfia curve. The women tagged subjects about 10 percent higher than the men; whereas, men tagged more descriptive terms, by 12 percent. In other words, the women tended to tag what the work was “about” and the men tagged “about the work.” The spikes in relational tagging for the men and affective for women were caused by the small sample size. Otherwise, the rest of the categories differed by one percentage point or less.

For levels of language specificity, only the genre and subject tags were examined, as the others do not have a comparable hierarchy. Genre was divided into two levels. Such terms as “Fiction,” “Literature,” or “Young adult” that gave no indication of “aboutness” were considered the broadest levels, with such tags as “horror” or “fairy tale,” making up the narrower level. The men tagged genre 60 times, with broad and narrow nearly split in half at 52 percent broad and 48 percent narrow. The women tagged genre 82 times. Of those, 36 percent were at the more specific level. Subjects were also divided into two hierarchical levels. The broader level included disciplines, aspects, or themes such as “Religion” or “Coming of Age.” The narrower level was made up of concrete terms such as “Potions,” “Cars” or “Zombies.” Of the men’s 41 subject tags, 36 percent was broad, whereas the women were split at exactly 50 percent broad to narrow. In this sample, men tended to tag genre equally broad and narrow, whereas women tended to be slightly more broad, which refutes Lakoff’s (1973) notion that women use more specific language. Combining subject and genre, men tagged more specifically 55 percent of the time, and women tagged more
specifically 47 percent of the time. Subject was evenly split by women, with men tending to be more specific. Only one misspelling was present (―autust‖ for ―August‖ by a man), which was not enough to support Coates’s (2004) assertion that women use more correct language. Overall, with the exception of the rates of tagging subject and descriptive tags and spikes caused by the small sample size, no dramatic differences were evident between men and women.

Discussion

In this limited sample, no discernible pattern of difference was evident between how men and women tagged. Users of both genders tagged prodigiously, minimally, broadly, and narrowly, calling into question the folklinguistic assumptions about how we use language. So does LibraryThing, as a community of practice, influence the way tags are composed? The community consists mostly of people who read, collect, and generally enjoy books, and a particular discourse associated with books seems to seep into the way users tag. Women and men both overwhelmingly created tag content that typically appears on a catalog record or commercial bookselling website: format, genre, year, general subject topics. Only 14 percent of tags for both genders were strictly personal. There is no way to tell if users examine other tags to shape their own practice, and in fact, the community of practice could be as much or more be about book discourse as about tagging practice. For example, the small percentage of affective tagging from this sample could be explained by an inability to evaluate the items, either because the items were still unread, because evaluation is not part of cataloging discourse, or because the site is public. Furthermore, LibraryThing offers a rating review and system, which provides users with a forum for opinions outside. Evaluative tags do exist. The Da Vinci Code alone garnered 46 instances of the tag ―crap‖, just not by this group of taggers. The ―empty adjectives‖ Lakoff describes were not evident for either gender.

Conclusion and Further Research

Based on the results of this exploratory study, gender has not shown to be an influential factor on how users tag. Tagging behavior in LibraryThing could be described using the community of practice framework; however, further exploration and a larger sample size are needed to identify to what extent the community has on linguistic variation in the online environment. The limitations of this study expose myriad opportunities to examine language use in social tags. A larger sample size could confirm conclusions from this study or some of the subcommunities within LibraryThing could be studied for language or membership requirements, such as the science fiction community. A broader discourse analysis could be conducted on reviews, collections, blogs, profiles and other texts on LibraryThing to discover more traits of the community of practice. Another intriguing avenue for further research involves investigating the community of practice in comparison to an epistemic domain to investigate the overlap between linguistic expression and meaning.

References


Eckert, Penelope and McConnell-Ginet, Sally. 2003. Language and Gender. New York: Cambridge UP.


Radia Bernaoui - ENSV of Algiers, Algeria
Mohamed Hassoun - ELICO Research Centre, France

User Expectations, Reality and Delineation of Agricultural Information Systems in the Maghreb

Abstract
In the knowledge society, collection, organization and transmission of knowledge are considered very important for the economic, technological and social development of a country. An information system for agriculture is being developed in Algeria based on an assessment of user needs. The proposed system will have FAO’s AGROVOC as the basis and will have multilingual features. This paper outlines some of the major requirements of the search interface and the vocabulary based on a survey of potential end users of the system.

Introduction
Information is a strategic resource and has a determining impact on the processes of decision-making, planning, management, and scientific research. Production, storage, diffusion, and exchange of information are major challenges for any organization. For this reason we speak now about “information society”. We see the emergence of a knowledge society within the information society. The purpose of what is called the “knowledge management” is to preserve, transmit and develop knowledge while increasing the intellectual capital. It acts both as a stimulant for innovation and as a factor of productivity. It also represents on the other hand, a device for enhancing the value of knowledge and competence transfer.

To develop a decision-making tool to evaluate the knowledge situation in certain subject areas we are developing an observatory /monitoring system / at the national level in Algeria. This requires the definition of "targets" of observation, and the type of information we are searching for, once the objectives of the observatory are clearly expressed. So, the primary application for the Local Information System (LIS) is to provide a location-centered reference base that is easily accessible to a wide range of users including data experts, managers, policy makers, front-line staff and citizens. They provide a wide range of statistics and reports allowing users to review the current evidence base and build a picture of localities and neighbourhood for their area of interest.

There are already some experiments to build observatories intended for this mission. In France, the Observatory of Sciences and Technology (OST) was created in order to conceive and to produce quantitative indicators related to scientific and technological activities and innovation (Généralités- Bibliométrie). An experiment supported by Canada allowed us to see how the observatory approach contributes to measure the evolution of science and technology at the national level (L’observateur).

We have been inspired by the experiments of neighbouring countries (Morocco and Tunisia). The Moroccan Institute of the Scientific and Technical Information is in charge of collection, organization and dissemination of scientific and technical information for the Moroccan and foreign scientific community (Moroccan Institute of the Scientific and Technical Information). In Tunisia, the National Observatory of Sciences and Technology of Tunisia has a mission of monitoring scientific and technological development. It is in charge of ensuring the follow-up of the progress as regards technologies and the realization of the prospective studies in the fields of development, research and innovation (National Observatory of Sciences and the Technology of Tunisia). Lastly, we present the Observatory of the National Documentary Potential (PDN) which can be considered as an
example of reflection on design of Algerian observatory. This is a diagnostic system for the scientific development of national libraries (Kasdi 1998). The objective of the presented project is thus to combine the various actions around a consortium to allow documentary resource sharing which will be standardized with the aim of installing a true national portal whose goal is to provide access to large databases. Information systems, in the complex universe, in which our corporations and scientists operate, must also function as management tools of the local resources of a country besides providing access to international resources. It is this double dimension of technological monitoring that will make our future information system a means of piloting research and development. Then, will emerge a collective intelligence characterized by the production of knowledge associated with the structuring and evaluation of collective actions, the construction of a memory, and the organization of individual competences, in spite of the geographical or cultural distances which separate the individuals concerned.

**National Observatory on Agronomic Research**

Algeria, faces the risk endangering its food safety (Bouzidi, 2008); it must therefore particularly consider the agricultural sector. In order to build an information system with added value to meet researchers’ expectations and adapted to their needs, a survey was undertaken among different research personnel. The study led to the creation of an information system, which will be accessible to the decision makers and scientists, a tool of piloting and increasing the value of the scientific research, by the development of several databases on the institutions, research laboratories, projects, the researchers and their publications. The query interface of the system which we propose will be presented like a Web page. This is being implemented by setting up a relational management information system for research purposes at the level of the Maghreb. It has several functions / features:

- **Two query and retrieval interfaces**: one for public access (Internet and Intranet), the other for restricted access designed for decision makers and research managers.
- A data management interface (access, update) allowing each institution to access its own data.
- An organization and input system presented as standard access forms for institutions, departments, laboratories, projects and individual researchers.

The total system will include and support:
- An interface to the current observatory database which will be used by the managers of the various research institutions.
- A database maintained by a pilot organization and managed by only one administrator.
- An annual safety copy for the conservation of the scientific inheritance of the various institutions.
- A history resulting from the annual safety copy, allowing the production of evaluation indicators of scientific research.

**Survey**

A large national survey was conducted for the purpose of knowing the degree of knowledge of scientists on different criteria and evaluation methods. A total of 500 questionnaires were distributed. The survey lasted five months (February - June 2008). We received 395 filled in questionnaires of which 345 were correctly filled (69%). The study was intended to evaluate the information needs of the scientists which should be available via the proposed Algerian information system of agronomic research. This analysis is based
on the needs expressed by the researchers concerning access to information in a variety of fields: national programmes of research, national and international publications, databases on the actors, tools of technological monitoring and collaboration.

**Access to National Databases**

Concerning the various databases planned for the information system to describe the national status of research in relation to the demands expressed by the scientists is quite balanced. The strong homogeneity of the responses shows that the researchers wish to have information on all the research details in its various components. This confirms the choice of the agencies of the observatory project which have planned to create five distributed federated databases: organizations, research laboratories, research projects, researchers, and research publications.

**Access to National Publications**

The analysis of researchers’ demands concerning access to national publications, according to the type of publication, indicates the information sources to be taken into consideration by the future information system on the agronomic research. The scientists express a clear preference for journal articles (25%), which communicate information at best scientific level, relevant and topical in the field of the agronomic research, either fundamental or applied. The theses, come in second position, with 20%; they are followed by lectures and books (18% for each of the two types). The reports represent 12% of the demands. On the other hand, the investigation reveals a lack of interest in national research patents for the development of innovations and the protection of knowledge, in spite of their interest to develop innovation and the economic potential of the country. This disinterest marks, for some, an ignorance of the existence of this type of publication in their field.

**Access to International Information**

Access to international information is a pre-requisite for scientific research innovations. The survey gives the following results: 24% of the expressed demands concern access to international electronic reviews, 21% the access to scientific events and 20% the retrieval from international databases. A result deserves a deeper consideration: open access represents only 12% of the demands. Exploratory interviews highlighted the ignorance of the scientists relating to this type of information resource (Open access and HAL - Hyper Article on-line). These results correspond to a study performed in 2007 on “The initiative of open access in Algeria” with a sample of 108 people: 78% of the Algerian researchers seem to be unaware of the existence of open access repositories (ArchiveSIC, arXiv and HAL; cf. Amrouni, 2007). Another study on “The electronic edition as tool for the enhancement of the agricultural scientific research in Algeria” made in the same year showed similar results: 80% of the researchers are unaware of the existence of open access (Bellahreche, 2007). We note that the researchers are unaware that open access sites facilitate access to publications, accelerate scientific exchange and improve consistency of the stored data. This collective initiative of the free access movement emphasizes the relevance of co-operation for sharing knowledge, production of innovations, and the creation of a knowledge society. In the same way, the tools of scientific monitoring do not seem to be known. The news accounts only for 8% of the demands and alerts (sd) 6%. These results affirm that, within the Algerian research institutions, there is not only non-existence of tools adapted to circulation, diffusion and division of scientific and technical information, but that there is also a lack of interest in these tools among researchers. Hence, the Algerian researcher is unaware that these tools currently constitute the principal means
of access to scientific information; that they enable knowledge sharing and facilitate the communication and the interactions with the centres of excellence.

**Access to Collaborative Platforms**

As collaborative work tools intended for scientific production and exchange, the Algerian scientists classify in the first position e-mail with 27% of the expressed demands. According to Poissonet (2002), “e-mail is based on a representation of exchanges as a space of a singular meeting between two subjects” Then, we have the forums of specialized exchanges with 20% of the demands. These are also tools which support the membership of virtual scientific communities and performing collective scientific productions. Among the tools for distance collaboration that the researchers wish to find in the information system, we have e-learning with 17% of the demands and the remote co-publication with 16%. The rate of these two media, according to the answers, does not correspond to the extent of means set up by the Algerian Ministry for Higher Education and Scientific Research (MESRS). Consequently, for the policy followed by this ministry to develop the e-learning in Algeria, it was agreed to create a facility for teacher training in the field of ICTs, the mobilization around new telecommunication technologies and teleprocessing contributing to the improvement of the quality of teaching, with a greater democratization of the access to the university. Indeed, the Algerian researchers work more in „bulk-heading” than in co-production. The exceptions apply to collective products made by researchers who are connected via networks associating external organizations or in collaborative networks within the same research department. The videoconference appears only in 14% of the demands. This application makes it possible to organize conferences between people who are separated by geographic distance and who do not need to move in order to establish distant contacts and exchanges. At last, the chat (asynchronous messages) which makes it possible to constitute living rooms of discussions is also of limited interest (15% of the demands). The scientists consider that the chat is rather a tool of leisure and distraction. It is important to specify that this type of exchange is usually used within the scientific social networks and can meet, initially, a personal need for discovering others on the Internet. This data reveals how much certain preconceptions can force the bulk-heading research activities and the isolation of researchers. Contrary to what the scientists think, this type of tool proposes also services of collaborative information management between researchers. It can be used in order to have relation and to create project teams.

**Agricultural Information System in the Maghreb**

The survey that we conducted among more than three hundred Algerian researchers reveals a fundamental need for the creation of a framework for the division and organization of knowledge to get more collective intelligence. The survey also helped identify a number of issues. Those issues that have a bearing on the theme of this conference will be highlighted in the following paragraphs.

With respect to future international projects, it is necessary to think of creating a Maghrebian information system devoted to the best processes of renewal and development of the mechanisms of research related to Maghrebian agricultural co-operation. The focus should be on the creation of an information system of high added value, which will be in the hands of the decision makers and the scientists, a tool of piloting and increasing the value of the scientific research, through the development of several databases on the institutions, the research laboratories, the projects, the researchers and their publications that can be effectively and efficiently searched. This implies an information system that has a common language for research exchange which means however, the application of multilingual features in the construction of the documentation language, e.g. in the sector of
agronomy. In particular we are thinking of the control of indexing and vocabulary (Richard 2011, Soergel 1974). The creation of a specialized information system on agronomic research in Maghreb is essential. Accordingly, it is possible to retrieve data of the information system whatever their nature, e.g. by a search in full text over all the headings of all the tables. Further features will be the selection of the required words and a possibility to limit the research objects. All key tables and words were developed starting from several thesauri such as “Agrovoc” (Agricultural information management standards) and “Agris” (International system of information for agricultural sciences and technology). Different hierarchical and associative relations (broader/narrower terms, related terms, equivalent terms, combination use) are established between the terms. It is suited for indexing and searching documents, web pages and digital objects. Agrovoc has also been used in combination with linked open data techniques to connect diverse vocabularies and to build the backbone of retrieval on Internet data. We think it’s very important to work out a standardized harmonized language and to build a thesaurus common to all the adherent organizations at the observatory. The tool of indexing of the thesaurus AGROVOC will facilitate, on the one hand, the indexing of the scientific production of the research institutions and the retrieval of the various relational databases of the Maghrebian observatory on the other hand. The advantages of the multilingual agricultural thesaurus AGROVOC are to represent more semantics than the specific relations between the concepts and the relations between their multilingual lexicalizations, for example, “a synonym”, “a translation”. It is a resource which structures and standardizes agricultural terminology in multiple languages, being able to be used by different users and systems worldwide.

References

1 AGROVOC has been created in the 1980. (AGROVOC 2012)


ABSTRACTS OF POSTERS
Semantic Web Technology: A Key for Effective Knowledge Management: A Case Study on Facets of Law

Abstract
Economic globalization, high speed internet and satellite supported telecommunication have added a new dimension to ‘Knowledge Management’ in terms of use of scholarly as well as business or commercial information. The qualitative and quantitative changes in the growth pattern of literature and information and changes in the knowledge seeking behaviour of specialized users throw up new challenges to professionals involved in the storage, management and effective retrieval of knowledge. The nature of research is also being affected compared to what it had been 10 or 15 years ago specially in the fields of science and technology and even in the social sciences. The legal issues associated with digital resources in terms of their online access and re-utilization have become very complex and sometimes confusing. To the librarian, the effective handling and management of these millions of pieces scholarly literature in digital form have become more complex. Besides, the inherent interdisciplinary character and complexities of digital rights management (DRM) adds to the problems. Law is an important and complex area seen from its applicability to varied disciplines. Due to the immense importance of law as a subject, and availability of plethora of related digital sources (i.e International & National Databases on Law), research and studies in this discipline have resulted in interdisciplinary and multidisciplinary scholarly literature with potential applications in a wide range of areas. The inter-operability issues associated with the large number of online databases and resources calls for proper metadata management for effective retrieval. Considering all these points, the facets of law and its penumbral areas need an effective management of metadata.

It’s true that the interoperability protocols (z 39.50) of the Internet facilitate access to databases or files from any domain and also permit export and import of data. However, from the standpoint of effective online information retrieval with good recall and precision, it has become more difficult and complex. The shortcoming of the relational matrix of the metadata platform of the digital library infrastructure is seen as the factor affecting effective information retrieval. The semantic web technology is considered in this study as a way to overcome the shortcoming of the metadata platform, where the metadata will be arranged in an ontological semantic manner in a classaurus like structure. Considering all these factors, an effort has been made to draw a prototype model where all the key facets / isolates of Law are mapped into a SKOS model (Simple Knowledge Organization System), where the facets of law will be in an ontological semantic matrix with relational hierarchy of facets. Here the ‘relational’ hierarchy means the thesaurus like structure of the facets, where they are arranged indicating relationships such as UT (Upper term), BT (Broader term), RT (Related term), and NT (Narrower term). And then the facets are mapped into SKOS graph and then those relational facets of the SKOS graph are being translated into RDF:XML program.
Knowledge Representation by Social identity Categories:
The Brazilian Institute for Web Science Research, A Case Study

Abstract
The task of creating conceptual structures to represent specific areas of knowledge has been a real challenge. Even with the opportunities offered by electronic programs such as specific software for the construction of conceptual maps and ontologies which help us in this task, an accurate representation of an interdisciplinary domain demands from the specialist a deep knowledge of its specific characteristics, its boundaries, and possible conceptual relations, in short, the true identity of the domain that is being represented. This study intends, firstly, to discuss a theoretical possibility of how an interdisciplinary field of knowledge can be represented by the categorization process derived, for example, from conceptual analysis theories. Secondly, it analyses the methodological principles by which an interdisciplinary area constructs its own identity representation. And thirdly, it studies the existence of an extensive and consistent application of the research results to other interdisciplinary fields of knowledge. To achieve these objectives the Brazilian Institute for Web Science Research (BIWSR) self representation discourse was considered as the empirical object of analysis. The theoretical foundations of concepts and categories in Dahlberg’s information model (1982) and in Fairclough’s critical discourse analysis (1989) were used to analyze the interdisciplinary domain of computer science from the point of view of BIWSR. The Brazilian Institute for Web Science Research was created due to the appearance, in 2006, of a new research domain – the so-called Web Science (Berners-Lee, 2006; Hendler, 2008). In this new domain, Web is the primary object of study and ceases to be viewed as a mere technology. As such, it involves not only research on computing and various technological aspects, but also on social and economic issues. More precisely, it is the science that investigates all issues around decentralized information systems, or unstructured information, covering people, software and hardware, and their multiple, complex interactions (Maculan & Lucena 2008). Using Fairclough's three-dimensional framework for studying discourse the analysis of discursive events as instances of socio-cultural practice is highlighted. And from the theoretical approach to categorize a domain the concepts/categories as knowledge units are highlighted (Dahlberg, 1992, p. 65-71). In this sense, the 5 main categories found in the BIWSR self representation were created considering: 1- People & Society, which investigates the social, political and economic aspects of the Web, based on the discourse used in government policies for participative and universal access to knowledge for the Brazilian citizen; 2- Software Technologies for Web Applications which considers specific issues in the design, development and deployment of large distributed applications on the Web, involving millions of users, based on the discourse of government policies for computational modeling of complex systems: artificial, natural, socio-cultural and human-nature interactions; 3- Management of Web
Data which addresses access and management of heterogeneous, distributed data sources, from the Terabyte (10^12), through the Petabyte (10^15), to the Exabyte (10^18) levels, based on the discourse of government policies for the management of information over massive volumes of distributed multimedia data; 4- Web Infrastructure which deals with the Web as a technological means to ensure scientific, technological and social progress, dealing primarily with the question of how to scale in order to meet performance or reliability expectations, contributing to all other categories; and, 5- Foundations of Web Science which is needed to support research in all of the previous categories, based on the discourse of government policies for technological development with quality: dependable, scalable and ubiquitous systems. The conclusions based on the aforementioned methodological approaches indicate that the five main social identity categories are mutually exclusive, complete in relation to the purpose of the domain, and irreducible. They represent the facets for which the domain is known and they have as literary warrant and user warrant government policies proposed for this specific area. The predication conceptual analysis also indicates the existence of a logical/semiological structure of the reference item (each research project), in relation to the categories e. g.: a project on the study of the process for the creation and preservation of digital heritage – projects – project groups – social networks – People & Society. Finally, it is concluded that it is possible to apply the methodology of this empirical experiment to other interdisciplinary areas, taking into account the theoretical studies of concept relations in addition to literary and social warrants found in the discourse of existing social practices for the knowledge domain that we want to represent.
Abstract

CATIA, Computer Aided Three-dimensional Interactive Application developed by Dassault Systèmes and marketed worldwide by IBM, CATIA V5 has been used by companies such as Airbus, The Boeing Company, Audi, Porsche, BMW and more. It is a commercial software used for Computer Aided Design (CAD), Computer Aided Engineering (CAE), and Computer Aided Manufacturing (CAM) in the aerospace and automotive sectors. The present study of the main CAD workbenches in CATIA V5 analyzes the main features for information visualization, information retrieval and Knowledge Organization. In addition, the study proposes a specification and application of the Dublin Core metadata scheme for the description of the main CAD objects in CATIA V5 that includes a simple way to defining and loading the scheme within the system and exporting the records without the help of any external applications or Application Programming Interface (API). Not many works studying such a specific topic of CATIA V5 have been published in the literature, and when it has been studied, it was usually done in talks about the ISO 10303-209 STEP AP standard and never about the application of Dublin Core. According to the published literature, the Dublin Core scheme has never been applied to such a specific area as CAD objects in CATIA V5. CATIA V5 is a powerful tool that can be easily customized via API's using Visual Basic or C++ programming languages. Following these possibilities, several commercial applications have been developed in order to enhance Product Lifecycle Management (PLM) purposes in CATIA V5 (E.g. Enovia and Windchill PDMlink). The present work analyzes the structure of the main workbenches used for CAD in CATIA V5, focusing on the PART design and Assembly design workbenches, and information visualization devices like the 3D space and the specification tree. As for the information retrieval features, possibilities in the advanced search are analysed, focusing on the characteristics of searchable fields of both Part design and Assembly design objects (i.e. .Product and .Part files), the specification of types of these fields that will be necessary for the definition of the scheme, and the possibilities the default search tool in CATIA V5 allows for these objects. Furthermore, an application of the fifteen Dublin Core elements, according to the Dublin Core Version 1.1 definition (http://dublincore.org/documents/dces/) for the context of CAD objects in CATIA V5 is discussed. Finally, an easy way to define all the Dublin Core elements within the system in two different simple ways is explained, using both the Bill of Materials and Listing Report CATIA V5 features, allowing an exportation of these records in .html, .xls, and .txt formats. The relevance of this work is in using the Dublin Core metadata scheme for the description of objects generated in CATIA V5 enabling automatic or semiautomatic processing of the exported data, thus allowing integration of these objects into broader collections, the establishment of relations with other objects of similar nature, and the comparison of these with other general objects of different nature, but capable of being described by the same scheme. The objectives of this description would also include the possibility of integration of CAD objects in digital repositories, academic learning environments, simple applications for designers, and any other system based on ontologies and recommender systems capable of inferring relations between objects. For this a general standardized scheme for representation and export of objects, as well as a definition capable of being applied to the CAD context in CATIA V5 is needed, and thus the specification of the Dublin Core scheme for this purpose justified.
Building the Ontology Dictionary using the Inference Engine RACER

Abstract

The term ontology is used with several different meanings. One of the first definitions was given by R. Neches and colleagues, who defined an ontology as follows: An ontology defines the basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary. This article aims to study the relationship of semantic and non-linguistic relationships in the ontology dictionary. Then we compare these relations with lexical relations occurring in the information languages used in the Polish library catalogs (for example: language subject headings of the National Library, language subject headings KABA). This goal is realized by building the ontology dictionary (conceptual schema) that contains concepts, the definitions of the concepts, and the specification of relationships among the concepts. Terms included in the subject index of Encyclopedic dictionary of information, language and information retrieval systems were used in the creation of ontologies. Some of these terms have been translated into English by using English-Polish dictionary of scientific information and librarianship.

The inference engine RACER (Renamed ABox and Concept Expression Reasoner) was used to build the ontology. The RACER implements TBox and ABox reasoner for the logic SHIQ. This paper discusses the basic functions of management knowledge base, which consists of terminological axioms (forming the so-called Tbox) and assertions (assertions forming the Abox). This ontology dictionary presents the ability to find words or phrases by means of functions and macros available in the engine RACER (see Fig. 1).

Fig. 1: Screenshot of the engine RACER: start-up TBox Semantyka

The logical relations defined in the description logic SHIQ enable semantic queries which can discover not just exact matches but logically related concepts. These relationships are similar to the relationships in information languages used in the Polish library catalogs (syntactic categories and grammatical categories). Analysis of the semantics of questions leads us to the conclusion that the inference engine RACER is an effective software for information classification, search and retrieval. The approach has been evaluated and can be used in studies of information languages in Poland.
Mehri Parirokh, - Ferdowsi University of Mashhad, Iran
Rahmatollah Fattahi - Ferdowsi University of Mashhad, Iran
Abdolrasoul Khosravi - Ferdowsi University of Mashhad, Iran

Google Suggest, Cognitive Load Theory, and Query Expansion: How Web Searchers Perceive Keyword Suggests in Google

Abstract
Web Search engines provide users with different facilities to improve retrieval. Keyword suggests for query expansion in Google is a new facility. However, users may find it difficult to use. The aim of this research is to investigate the extent and the types of cognitive load (Extraneous, Intrinsic, and/or Germane) that the searchers encounter in using Google suggested keywords. Little research has been reported in the area of using suggested keywords in the process of query expansion. The methodology of this research helps access to valuable results in this regard.

Design/Methodology/Approach: Through a mixed approach (using a questionnaire as well as “think aloud” technique) quantitative and qualitative data were collected from 60 post-graduate students from two broad fields of Humanities and Basic Sciences at Ferdowsi University of Mashhad, Iran.

Findings- Findings show that suggested keywords and retrieved pages did not make any significant problem (or negative cognitive load) for users. Also, there was no significant difference between Humanities students and Basic Science students regarding cognitive load when they were browsing the suggested keywords. Most of the suggested keywords and retrieved pages could fulfill users’ needs and had direct relationship with their primary keywords.

Conclusions- Suggested keywords/queries can help Web searchers learn better and expand their queries faster. On the other hand, based on users' reactions in this research towards the list of suggested keywords, designers could improve the search interface and provide a more relevant structure for optimizing queries.
Marc Tanti - University of Aix-Marseille, Marseille Cedex, France

“Naming” Electronic Documents: Development and Evaluation of a Procedure

Aim and Scope of the Study
The documents are ubiquitous in everyday life. However, the transition to digital has completely altered the overall context - ways of access and sustainability. The issue of "naming" the document at the time of registration in the memories of computers is a recurring issue. Indeed, what name to assign to enable better sustainability? Better accessibility? Better identification? Better interoperability in different environments? These issues are particularly important because this "naming" can be a great asset to index and search the information. After a theoretical analysis of the issue, this article proposes a procedure of "naming" the document. The procedure is also evaluated to determine its utility in information retrieval.

Theoretical Analysis of the Issue
A search was performed on the databases of AFNOR (Association Française de Normalisation), IOS (International Organization for Standardization), WOS (Web Of Science) and on Google® with the following equations: "Filename", "Filename" AND "Rules", "Filename" AND "Standard", "Filename" AND "Nomenclature", "Filename" AND "Good practices", "Filename" AND "Best Practices". None of the documents found described a “naming” standard or procedure. However, 15 significant documents were selected, including "The Standard Naming Conventions for Electronic Records" (Thompson, 2007), published by Edinburgh University which advocates a few rules. We will adapt some of these rules to propose a “naming” procedure that will allow better indexing; we offer 18 rules.

Naming Procedure
We propose to give to each document a simple, meaningful, orderly and understandable (25-64 characters) file name. We propose to avoid areas commonly translated as "20%" by search engines. We recommend using a 2-digit code if a number is included in the document name and YYYY, if it's a date. We recommend writing the first name if a person is included in the document name. We propose to avoid "empty words" or common names, to avoid using the asterisk, the sharp, the accents, the ampersand (&), quotes (French or English), exclamation points, question marks and ellipses, the signs of operation (+, -, *, /), the vertical bar, the signs of comparison (<, >, =), brackets, diacritical marks, the space, ", #, %, {, }, \, ^, ~, [], `;, /, ?,:, @, =, &). We recommend avoiding plurals, not to separate the letters with dots for acronyms. On these rules, we propose to name the documents by a general keyword and 2-3 specific keywords that best represent the content of the document, followed by the first author and the year of publication. Keywords, author and year are spaced by a hyphen and are written in small print ("generic term-specific terms-first author-year"). The generic term consists of a single keyword, assigned by the indexer, based on the title, subtitle, introduction and conclusion of the document, in natural language or using a thesaurus. Up to three specific terms could be assigned. For example, a document dealing with the diagnosis of malaria written by Dubois in 2006 will be named "malaria-diagnosis-dubois-2006".
Evaluation of the Procedure

This procedure was applied to a corpus of 12,000 documents from the databases MedLine, Horizon, BDSP (Base de Données de Santé Publique), WOS (Web of Science), ScienceDirect, Wiley, EM-Consult. This corpus is integrated into a software of Electronic Document Management (EDM) (Gargantua®). The documents deal with various health risks for the military in operation (malaria, chikungunya…) and are accessible via an intranet to 112 users (military doctors, pharmacists, veterinarians…). A questionnaire for assessing the relevance of the names of documents for sanitary information search was distributed to 10 users randomly. 100% responded that this procedure allowed them to find information more easily and faster. 100% found that names assigned were neither too long nor too complicated. 20% expressed some reservations about some key words in the form of acronyms. 100% responded that the type of name given to documents brought information about the content without having to open and read it. 60% said that the first author's name in the name of the document may not be much help except in cases where an author search was conducted. 100% found the year of publication in the name of the document useful to sort the documents by date. 100% suggested that the classification of the documents within thematic files brought implicit information about the content of the document. Finally, 70% said more keywords in the name of the document would improve the ease and speed of research.

Conclusion

Our study demonstrates the absence of standard for “naming” documents. However, the University of Edinburgh published "The Standard Naming Conventions for Electronic Records" (Thompson, 2007) containing rules that we propose to adapt in a procedure for “naming” documents. This procedure was applied to a corpus of 12,000 documents about health risk in operation. A panel of 10 medical officers evaluated the result. 100% of respondents said that this procedure allowed them to find information more easily and quickly. In addition, it brought information about the content without having to open and read it. It decreased the time needed to find information. The procedure needs to be evaluated by all users of the corpus (112). It is also necessary to test the procedure on documents dealing with other topics and other users.

Bibliography

Beaulieu P. 2004, Eléments à prendre en compte lors du nommage de fichiers et de dossiers, INU1020 Organisation de l'information numérique, EBSI, Université de Montréal.

Department of Cultural Resources. 2008, Best practices for File-Naming, Office of Archives and History Government Records Branch.


Tremblay, M. 2008. Les 3 règles d'or pour nommer un fichier, Hors des lieux communs.
Aspects of Information Organization and Retrieval from a News Portal

Abstract

It is a challenge to organize the contents of a news portal, in which the most evident aspect is the encyclopedic character of the information produced and aired daily. Usually, the users of such services increasingly use the portal as a source of daily information, replacing printed media. In this context, news production includes not only its preparation itself, but also the creation of significant links between texts, images and sounds. Therefore, in this environment, the journalist/editor simultaneously assumes the roles as portal’s producer, organizer and user. The process of creating news in an electronic portal leads to dispersion of related issues and to improper relations between themes for different reasons: intense work pace, significant number of professionals involved, lack of policies and shared devices to treat and retrieve information, among others. These issues cause breaks in. The relation between adequate and consistent subjects is, thus, an important factor to retain users. This is the problem that motivated the development of this research. Theoretical and methodological aspects of the research are presented here, as well as a set of parameters of creation and maintenance of controlled vocabularies. It is also presented a manual of good practices of indexing pages in the portal. The study was done based on: Domain Analysis and Guidelines for the construction of languages knowledge Organization. The empirical study was conducted at the news portal UOL (Universo Online), a company that has a staff of more than 200 journalists, editors and publishers, that has an audience of more than 29.8 million visitors and more than 4.214 billion page views per month. It currently has 2.5 million paying subscribers for its services to access content and products. (http://sobreuol.noticias.uol.com.br/, 20/10/2011).

Theoretical and Methodological References

Domain analysis was proposed in the 1990s to study domains of knowledge . (Hjørland; Albrechtsen, 1995 p.400). From this perspective, we should not treat the knowledge fields as if they were all fundamentally similar. In Domain analysis, several applications are proposed (Hjørland, 2002):
1. Producing literature guides or subject access;
2. Construction of special classifications and thesauri;
3. Specificity and specialty of indexing and information retrieval;
4. Empirical users studies involving investigation of the information needs of different communities.
5. Bibliometric studies, to analyze links between documents.
6. Historical studies, considered as fundamental methods to provide deeper perspectives and relevant context of the studied domain.
7. Documents and genres studies, to support studies of Information Architecture.
8. Epistemological and critical studies, to substantiate critical analysis of the knowledge domains.
9. **Terminological studies, special purposes languages (SPL), database semantic and speeches studies.**

10. **Structures and institutions in scientific communication** to support the characteristics, to identify the actors and map the information flows in this domain. The structures studies of the internal divisions inside the working areas and all the information exchange between domains provides useful information for understanding functions of specific types of documents and information services.

11. **Scientific cognition expertise and knowledge about artificial intelligence.** This approach can provide useful techniques to supplement other approaches to domain analysis in information science. As noted by the author, Domain Analysis has long been used in computer science, but for other purposes. Even so, relies on useful techniques for our area, which should be adapted.

In this research information and language issues are considered as part of a cultural system that influences information organization and retrieval (Hjørland and Albrechtsen, 1995, p. 407). In this sense, taxonomy and folksonomy or social information classification concepts were discussed. Quintarelli (2005) defines folksonomy as a new approach to classification of distributed digital resources. Likewise, it is seen as an approach to classification (Hammond, 2005). For these authors, it is an unstructured classification made by digital resources users. Can the publishing systems for news portals be considered social tools? While its use is collective, is there interaction? The only feature that could bring the language of the portal system close to folksonomy is to use natural language to create and assign descriptors to each tag which, according to Lancaster (1993), is the language commonly used in writing and speech, and that is the opposite of ‘controlled vocabulary’. For semantic and cognitive aspects of classification, Golder and Huberman (2006) say that labeling systems need to attack problems that are inherent in the process of creating semantic relationships between concepts. To create the controlled vocabulary for the portal, the three main problems of language - polysemy, synonymy and basic level variation - were considered.

**Results and Provisional Conclusions**

To solve the mentioned problems we considered the need of a moderating (mediating) process for creating terms, which was based on methodological theories on information and knowledge organization and guidelines such as ANSI / NISOZ39.19 (2005) of creation and maintenance of controlled vocabularies. The following conclusions resulted: a) information organization and retrieval from portals requires a combination of automatic processing and indexing tools, given the quantity and diversity of news published each day; b) it is necessary that the tool has clear syntax, easy understanding and fast processing, to any Java-based application. The tool should allow the management of texts and images publication (photos and infographics) and the integration of text and related images, in the form of an album and finally, a set of good indexing and updating of the system vocabulary.

**References**


K. S. Chudamani – Indian Institute of Science, Bangalore
H. C. Nagarathna - Indian Institute of Science, Bangalore
B. Sharatha - Indian Institute of Science, Bangalore

Clouds and Libraries:
Classification as a Tool for Cloud Computing

Abstract:
Cloud computing is a new technology model for IT services which many businesses and organizations are adopting. Sharing of resources is a common component of work in libraries, especially with networking. When it is in a cloud, there is a common development platform. It comprises of both hardware and software, which is reusable, and constant updates indexes the online catalogue. In Libraries, cloud computing can simplify processes, save time and money. This article defines cloud computing using classification and shows how it is different. It also discusses how cloud computing solutions using classification systems could be beneficial to libraries.

Subjects, authors, title, etc form clouds, which can be subdivided into sub-clouds, while searching on a computer. Even if catalogues of different libraries using the same classification are placed in a remote source, common searching is possible. How to handle the cloud is like any other index, except that one has to choose a particular cloud. Once a cloud is chosen, normal operations such as moving the highlighting bar to the corresponding entry are available to display bibliographic details. Cloud is at the heart of the retrieval system organized in the order of class number. Each class number along with its subdivisions forms a cloud and sub-clouds. The whole classified system is like a grid in which clouds and sub-clouds are embedded into it. Entering a class number retrieves a class cloud and its sub-clouds. First three characters of book number may also be given to position the classified cloud more accurately.

All clouds are in one way or the other classified. The author, title, etc are alphabetically classified. The keyword and the Boolean and other related operators are also categorized as belonging to a cloud, which is an intersection cloud. But, the most important cloud is the classified, as it enables browsing. An examination of the classified cloud helps us in understanding the deeper ramification of subject clouds. The metadata required for the purpose of creating such clouds is quite complex. First of all, super ordinate and subordinate relation needs to be defined. The example of Physics is used here to illustrate the concept. Physics is a big cloud or can be called “PHYSICS Grid” including its canonical clouds. Canonical clouds are formed on the basis of a general understanding of the extension of the subject over generations. Formation of clouds - such as compound, complex and intersection clouds - varies with the classification scheme. For example, in Colon Classification physics and engineering are placed adjacent to one another. In DDC they are far away. In-between, chemistry, biology, earth science, etc are located. Accordingly, the representation varies while searching using class numbers.
The Efficacy of Google's Suggested Keywords and Phrases in Query Expansion Based on "The Least effort Principle" and "Cognitive Load Theory"

Abstract
Purpose: One of the most effective ways to improve relevance performance in IR systems including Web search engines is to provide their users with tools facilitating query expansion. Search engines such as Google provide users with keyword suggest tools. The aim of this research was to investigate Users reactions regarding the efficacy of Google's suggested keywords/queries in query expansion based on "The Least effort Principle" and "Cognitive load Theory" from users' view

Design/Methodology/Approach: Through a mixed method approach quantitative and qualitative data were collected from 60 postgraduate students at Ferdowsi University of Mashhad, Iran (Humanities & social Science - Basic and engineering science) using four different instruments (questionnaire, thinking aloud technique, query logs, and interviews).

Findings: Among others the ‘relation between suggested keywords and the information need’ (mean rate of 3.53 out of four) was considered the most important by searchers in selecting suggested keywords for query expansion. Also the ‘relation between suggested Keywords and the retrieved items’ (mean rate of 3.62) was considered the second most important criterion in judging the relevance of the retrieved results. The participants agreed that the suggested keywords improved relevance.

According to the type and the level of suggested keywords/queries cognitive load, the results show that suggested keywords/queries and retrieved documents have not made any negative cognitive load for uses. Most of the suggested keywords/queries and retrieved pages have conformity with users’ needs. The other hypothesis test show that there is a significant difference between the relevancy of retrieved results of primary queries and retrieved results of query expansion based on suggested keywords/queries. In other words, suggested keywords/queries help improve the relevancy.

Based on “the least effort principle”, findings show that over 87% users considered and used Google’s suggested keywords/queries. There is no significant difference between the usage of suggested keywords/queries among different groups of users - liberal arts and the other fields of sciences. Finally some applicable solutions for improving the relevancy of retrieved suggested keywords/queries have been suggested.

Originality/value: This research makes a contribution to the need of designers of IR systems to support query expansion. Another contribution of the study is the identification of a number of new relevance judgment criteria for web-based environments.
Umesh Kulkarni - Y. M. College, Dept. of Library and Information Science, Pune, India

V. S. Mugade - Y. M. College, Dept. of Library and Information Science, Pune, India

N. B. Dahibhate - National Chemical Laboratory, Pune, India

M. P. Wadekar - Y. M. College, Dept. of Library and Information Science, Pune, India

Vividha Dhapte - Poona College of Pharmacy, Pune, India

B. A. Kulkarni - Bharati Vidyapeeth, Pune, India

S. S. Kadam - Bharati Vidyapeeth, Pune, India

Organization of Collections of Postal Stamps

Introduction
Postal stamps are significant artifacts in our daily life as also in the history of a nation. Due to their attractive and intellectual designs these postal stamps are the subjects of special interest. Thousands of stamps are published annually all over the world. Along with these stamps usually related literature giving the purpose, background and importance of the stamp and the occasion on which it was published is issued. A notable aspect of stamps is that they cover the entire spectrum of knowledge. This paper is based on an ongoing project to develop a database of stamps; an effort has been made to apply KO techniques for organizing these.

Classification of Stamps
The most convenient and logical way for organizing postal stamps is to classify them into different disciplines of knowledge. Researchers appear to prefer discipline-based classification. Each discipline may be further subdivided.

There are innumerable postal stamps published up till now. A limited and representative sample of stamps was selected for this experiment in the present project. About one hundred (100) postal stamps representing multiple disciplines and aspects were selected.

Table 1: Discipline wise classification of postal stamps for KO

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>Name of Discipline</th>
<th>Related faculty</th>
<th>Name of Stamp</th>
<th>Related Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Arts</td>
<td>Arts</td>
<td>Rabindranath Tagore</td>
<td>India</td>
</tr>
<tr>
<td>2.</td>
<td>Sciences</td>
<td>Science</td>
<td>Radio Century</td>
<td>Italy</td>
</tr>
<tr>
<td>3.</td>
<td>Management Sciences</td>
<td>Commerce</td>
<td>J. D. Birla</td>
<td>India</td>
</tr>
<tr>
<td>4.</td>
<td>Social Sciences</td>
<td>Metal, Moral</td>
<td>Mother Teresa</td>
<td>India</td>
</tr>
<tr>
<td>5.</td>
<td>Mental and moral sciences</td>
<td>And Soc. Sciences</td>
<td>Swami Vivekanand</td>
<td>India</td>
</tr>
<tr>
<td>6.</td>
<td>Medical Sciences</td>
<td>Medicine</td>
<td>Physiology/ Medicine</td>
<td>Grenada</td>
</tr>
<tr>
<td>7.</td>
<td>Engineering Sciences</td>
<td>Engineering</td>
<td>Sir Vishveshwarya</td>
<td>India</td>
</tr>
<tr>
<td>8.</td>
<td>Educational Sciences</td>
<td>Education</td>
<td>Mahatma Phale</td>
<td>India</td>
</tr>
<tr>
<td>9.</td>
<td>Pharmaceutical Sciences</td>
<td>Pharmacy</td>
<td>Dr. Hansen</td>
<td>India</td>
</tr>
<tr>
<td>10.</td>
<td>Agricultural Sciences</td>
<td>Agriculture</td>
<td>Rice Cultivation</td>
<td>Ceylon</td>
</tr>
<tr>
<td>11.</td>
<td>Environmental Sciences</td>
<td>Science</td>
<td>Tulsi</td>
<td>Nepal</td>
</tr>
<tr>
<td>12.</td>
<td>Space Sciences</td>
<td>Science</td>
<td>Space Achievement</td>
<td>U.S.A.</td>
</tr>
<tr>
<td>13.</td>
<td>Earth Sciences</td>
<td>Science</td>
<td>Comet</td>
<td>U.K.</td>
</tr>
<tr>
<td>14.</td>
<td>Other Subjects</td>
<td>Law</td>
<td>Dr. B. R. Ambedkar</td>
<td>India</td>
</tr>
</tbody>
</table>
The results suggest that classification schemes can be applied for organization of stamps and to support their retrieval. However, there is a need for projects to develop searchable databases of stamps; more work is on to develop multifaceted in depth databases.
LIST OF CONTRIBUTORS

Amaral, Roniberto Morato
UFSCar
Rua Oziel de Camargo Lima, 50
Jardim Acapulco. CEP 13563-701
São Carlos-SP, Brazil
roniberto@nit.ufscar.br

Arastoopoor, Sholeh.
Regional Information Center for Science and Technology, Shiraz, Iran
shole.arastoopoor@gmail.com

Arboit, Aline Elis
São Paulo State University – UNESP
737, Hygino Muzzi Filho Avenue,
17525-900 Marilia - SP – Brazil
aarboit@yahoo.com.br

Asundi, A.Y.
77, 3rd Main Road, Prashanth Nagar
Bangalore 560 079, India
ashok_asundi@yahoocom

Ávila, Daniel Martinez
Universidad Carlos III de Madrid
C/ Madrid 126, Getafe
Madrid) 28903Spain
dmartine@bib.uc3m.es

Babik, Wieslaw
Institute of Information and Library Science
Jagiellonian University in Cracow
PL 30-348 Kraków
ul. St. Lojasiewicza 4 Poland
w.babik@uj.edu.pl

Barros, Thiago Henrique Bragato
Sao Paulo State University - UNESP
Campus de Marilia, Brazil
sean.vogel@gmail.com

Bernaoui, Radia
Division of Information Sciences
Higher National Veterinary School of Algiers, Algeria
radiabernaoui@yahoo.fr

Bhanumurthy, K.
Scientific Information Resource Division,
Bhabha Atomic Research Centre,
Trombay Mumbai 400 085
headsird@barc.gov.in

Bufrem, Leilah Santiago
State University – UNESP, Brazil
santiagobufrem@gmail.com

Campbell, Grant D.
School of Information and Media Studies
University of Western Ontario
London, Ontario, N6A 5B7, Canada
gcampbel@uwo.ca

Cerdeira, Pablo de Camargo
Law School - Getulio Vargas Foundation
Rio de Janeiro, Brazil
pablo.cerdeira@gmail.com

Chada, Daniel Magalhães
Business Administration School Getulio Vargas Foundation – Praia de Botafogo
190 Rio de Janeiro, RJ, Brazil
daniellec2112@gmail.com

Chen, Hsueh-hua
Graduate Institute of Library and Information Science
National Taiwan University
No.1, Sec.4, Roosevelt Road
Taipei 10617, Taiwan
sherry@ntu.edu.tw

Chen, Shu-juin
CITI, Academia Sinica
128 Academia Road, Section 2, Nankang
Taipei 11529, Taiwan
sophy@sinica.edu.tw

Chudamani, K. S.
Indian Institute of Science
Bangalore 560 012, India
ksc@library.iisc.ernet.in
Codeço, Claudia Torres  
Center for Quantitative Methods  
Fundacao Oswaldo Cruz/PROCC  
Avenida Brasil 4365 Residencia Oficial  
Rio de Janeiro 21045-900, RJ, Brazil  
claudia.codeco@gmail.com

Coelho, Flávio Codeço  
Applied Mathematics School  
Getúlio Vargas Foundation– Praia de Botafogo, 190, Rio de Janeiro, RJ, Brazil  
fccoelho@gmail.com

Cordeiro, Rosa Inês de Novais  
Federal University of Fluminense  
Niterói, Brazil  
rosacordeiro@vm.uff.br

Corrêa, Carlos Alberto  
São Paulo State University, (UNESP)  
(Universidade de São Paulo (USP))  
Rua Prof. Lucio Martins Rodrigues, 443 São Paulo - SP – Brazil 05508-900  
calbcorreia@yahoo.com.br

D’Ambrosio, Amber  
University at Albany  
State University of New York, USA

Dahibhate, N.B.  
National Chemical Laboratory  
Pashan Road, Pune 411 008, India

Dasgupta, Abhijit  
Indian Association for the Cultivation of Science (IACS)  
(An Autonomous Body funded by DST, GOI)  
Jadavpur, Kolkata 700032  
library@iacs.res.in  
dasgupta.abhijit60@gmail.com

Dayyani, Mohammad Hosien  
Ferdowsi University of Mashhad, Iran  
daneshvarz85@gmail.com

Dhapte, Vividha  
Poona College of Pharmacy,  
Paud Road Pune 411 038, India  
vdhapte@rediffmail.com

Deokattey, Sangeeta  
Scientific Information Resource Division  
Bhabha Atomic Research Centre  
Trombay Mumbai 400 085, India  
sangeetadeokattey@hotmail.com

Dixit, D.K.  
Computer Division  
Bhabha Atomic Research Centre  
Trombay, Mumbai 400 085  
dkdixit@magnum.barc.ernet.in

Dodebei, Vera  
Federal University of the State of Rio de Janeiro – UNIRIO, Rua Assunção, 159 401B –Botafogo. CEP. 22251-030  
Rio de Janeiro, RJ - Brazil  
dodebei@gmail.com.

El Hadi, Widad Mustafa  
University of Lille 3- Laboratoire GERIICO- France  
widad.mustafa@univ-lille3.fr

Fan-Hua, Nan  
National Taiwan Ocean University,  
Taiwan  
fhnan@mail.ntou.edu.tw

Faria, Leandro Innocentini Lopes de  
UFSCar, Rod. Washington Luis,  
Km 235 Monjolinho, 13565-905  
Sao Carlos, SP – Brazil  
leandro@nit.ufscar.br

Fattahi, Rahmatollah  
Ferdowsi University of Mashhad, Iran  
fattahirahmat@gmail.com

Favier, Laurence  
University of Lille 3- Laboratoire GERIICO- France  
Laurence.favier@univ-lille3.fr
Fox, Melodie
School of Information Studies
IOrg Research Group
University of Wisconsin-Milwaukee
Milwaukee, USA
msjoblad@uwm.edu

Gao, Cheng
e-Humanities Group, Royal Netherlands Academy of Arts and Sciences – KNAW
Joan Muyskenweg 25, 1096 CJ Amsterdam, Netherlands
gaocheng.bupt@gmail.com

Gontijo, Miriam
Federal University of the State of Rio de Janeiro – UNIRIO
Rua Almirante Tamandaré 41
aptº 116- Flamengo- CEP 22210-060
Rio de Janeiro, Brazil
bhmgonjijo@ig.com.br

Gou-Chi, Lee
National Ilan University, Taiwan
kelee@niu.edu.tw

Gracio, Maria Cláudia Cabrini
São Paulo State University – UNESP
Marilia 17525-900, Brazil
cabrini@marilia.unesp.br

Gracioso, Luciana de Souza
Course Coordinator,
Library and Information Science
UFSCar, Rua Miguel Pebron, 4900, C.45, Sao Carlos
SP 16563-470, Brazil
luciana@ufscar.br

Guastavino, Catherine
Graduate School of Library and Information Studies
McGill University
3661 Peel, Montreal Quebec
H3A 1X1, Canada
catherine.guastavino@mcmill.ca

Guimarães, José Augusto Chaves
São Paulo State University -UNESP
Av. Hygino Muzzi Filho 737
Marilia - SP – Brazil
guima@marilia.unesp.br

Hajibayova, L.
School of Library and Information Science,
Indiana University Bloomington
1320 East 10th Street, LI 011
Bloomington, IN 47405-3907, USA
lhajibay@indiana.edu

Hassoun, Mohamed
ENSSIB de Lyon
ELICO Research Centre France
mohamed.hassoun@enssib.fr

Henttonen, Pekka
School of Information Sciences
FI-33014 University of Tampere,
Finland
pekka.henttonen@uta.fi

Higuchi, Suemi
School of Applied Mathematics – CPDOC
Getulio Vargas Foundation Rio de Janeiro Brazil
Suemi.Higuchi@fgv.br

Hjørland, Birger
Royal School of Library and Information Science
6 Birketinget, DK-2300
Copenhagen, Denmark
bh@iva.dk

Hoffmann, Wanda Aparecida Machado
UFSCar, Rua Norman Abbud
214 - São Carlos
SP CEP: 13563 703, Brazil
wanda@nit.ufscar.br

Iyer, Hemalata
University at Albany
State University of New York, USA
hiyer@albany.edu
<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
<th>Address</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jacob, E. K.</td>
<td>School of Library and Information Science Indiana University</td>
<td>Wells Library, 1320 East 10th Street Bloomington, IN 47405-3907, USA</td>
<td><a href="mailto:ejacob@indiana.edu">ejacob@indiana.edu</a></td>
</tr>
<tr>
<td>Julien, C. A.</td>
<td>Graduate School of Library and Information Studies McGill University</td>
<td>3661 Peel Montreal Quebec, H3A 1X1, Canada</td>
<td><a href="mailto:charles.julien@mail.mcgill.ca">charles.julien@mail.mcgill.ca</a></td>
</tr>
<tr>
<td>Kadam, S.S.</td>
<td>Bharati Vidyapeeth Deemed University LBS Road, Pune 411 030, India</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Karanth, P.</td>
<td>PES Institute of Technology, 100 ft Ring Road, BSK III Stage Bangalore 560085, India</td>
<td><a href="mailto:karanthpallavi@gmail.com">karanthpallavi@gmail.com</a></td>
<td></td>
</tr>
<tr>
<td>Khoo, C.S.G.</td>
<td>Wee Kim Wee School of Communication and Information Nanyang Technological University 31 Nanyang Link Singapore 637718</td>
<td><a href="mailto:chrishkoo@pmail.ntu.edu.sg">chrishkoo@pmail.ntu.edu.sg</a></td>
<td></td>
</tr>
<tr>
<td>Khosravi, A.</td>
<td>Bushehr University of Medical Science Bushehr, Iran</td>
<td><a href="mailto:khosravi2422@gmail.com">khosravi2422@gmail.com</a></td>
<td></td>
</tr>
<tr>
<td>Kobashi, Y.</td>
<td>Escola de Comunicações e ARtes - USP Av. Prof. Lúcio Martins Rodrigues 44305508-900 São Paulo - SP – Brazil</td>
<td><a href="mailto:nairkobashi@gmail.com">nairkobashi@gmail.com</a></td>
<td></td>
</tr>
<tr>
<td>Krishnamurthy, M.</td>
<td>DRTC, Indian Statistical Institute 8th Mile, Mysore Road 560 059, India</td>
<td><a href="mailto:mkrishna_murthy@hotmail.com">mkrishna_murthy@hotmail.com</a></td>
<td></td>
</tr>
<tr>
<td>Kulkarni, B.A.</td>
<td>Bharati Vidyapeeth, Deemed University LBS Road, Pune 411 030, India</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kulkarni, U.</td>
<td>Y. M. College, Pune 411 038, India</td>
<td><a href="mailto:kulkarniumesh25@yahoo.com">kulkarniumesh25@yahoo.com</a></td>
<td></td>
</tr>
<tr>
<td>La Barre, K.</td>
<td>Graduate School of Library and Information Science University of Illinois at Urbana-Champaign 501 E. Daniel Street MC-493, Champaign IL 61820-6211 USA</td>
<td><a href="mailto:klabarre@illinois.edu">klabarre@illinois.edu</a></td>
<td></td>
</tr>
<tr>
<td>Leide, J. E.</td>
<td>Graduate School of Library and Information Studies McGill University</td>
<td>Montreal H3A 1X1, Canada</td>
<td><a href="mailto:john.leide@mcgill.ca">john.leide@mcgill.ca</a></td>
</tr>
<tr>
<td>Lima, G. A.</td>
<td>School of Information Science (ECI) Federal University of Minas Gerais (UFMG) Av. Antônio Carlos 6627 Sala: 4053 – Pampulha 31270-010 Belo Horizonte/MG, Brazil</td>
<td><a href="mailto:glima@eci.ufmg.br">glima@eci.ufmg.br</a></td>
<td></td>
</tr>
<tr>
<td>Lorenzon, E.</td>
<td>Faculdade De Tecnologia De Jahu - Fatec Jahu, S.N., Frei Galvão Street 17212-599 Jaú, São Paulo, Brazil</td>
<td><a href="mailto:eilenalorenzon@gmail.com">eilenalorenzon@gmail.com</a></td>
<td></td>
</tr>
<tr>
<td>Maculan, B.</td>
<td>School of Information Science Federal University of Minas Gerais (UFMG) Rua Adolfo Pereira, 431/302 303130-350- Belo Horizonte/MG, Brazil</td>
<td><a href="mailto:benildes@gmail.com">benildes@gmail.com</a></td>
<td></td>
</tr>
</tbody>
</table>
Mahesh, Kavi  
PES Institute of Technology  
100 ft Ring Road, BSK III Stage  
Bangalore 560085, India  
drkvimahesh@gmail.com

Milonas, Elizabeth  
NYC College of Technology  
CUNY, New York, USA  
emilonas@CityTech.Cuny.edu

Ming-Shu, Yuan  
Shin Hsin University, Taiwan  
juanems@cc.shu.edu.tw

Mitchell, Joan S.  
OCLC, 6565 Kilgour Place  
Dublin, OH 43017-3395, USA  
mitchelj@oclc.org

Moraes, João Batista Ernesto de  
Faculty of Philosophy and Sciences  
FFC - São Paulo State University  
UNESP  
Avenida Higyno Muzzi Filho  
737 - 17525-900 Marília – SP, Brazil  
prof.joao@gmail.com

Mugade, V.S.  
Y. M. College, Pune 411 038, India

Murguia, Eduardo Ismael  
Information Science  
Universidade Federal Fluminense  
Rua Praia do Flamengo, 72  
Ap. 928 Flamengo. CEP 22210-030  
Rio de Janeiro, RJ Brazil  
murguia@vm.uff.br

Nagarathna, H. C.  
Indian Institute of Science  
Bangalore 560 012, India

Neelameghan, A.  
Sarada Ranganathan Endowment for Library Science  
702, 42<sup>nd</sup> Cross, 3<sup>rd</sup> Block, Rajajinagar  
Bangalore 560 010, India  
anm2002@vsnl.net

Nikam, Khaiser  
University of Mysore  
Mysore 570006, India  
khaiser.nikam6@gmail.com

Nizam, Naureen  
Dalhousie University, Halifax, Canada  
nizam@cs.dal.ca

Ohly, Peter  
ISKO, Prinzenstr. 179  
D-53175 Bonn, Germany  
peter.ohly@gmx.de

Oliveira, Ely Francina Tannuri de  
São Paulo State University – UNESP  
Marilia 17525-900, SP, Brazil  
etannuri@marilia.unesp.br

Orrico, Evelyn  
CEP.: 22430-190  
Rio de Janeiro, RJ – Brazil  
evelynorrico@uol.com.br

Padmavathi, T.  
Central Food Technological Research Institute, Mysore 57 0012, India  
Email: padmavathit@yahoo.com

Parirokh, Mehri  
Ferdowsi University of Mashhad, Iran  
mparirokh@gmail.com

Pattuelli, Maria Cristina  
School of Information and Library Science Pratt Institute  
144 West 14th Street, New York  
NY 10011-7301, USA  
mcpattuelli@gmail.com

Raghavan, K.S.  
DRTC, Indian Statistical Institute  
8<sup>th</sup> Mile, Mysore Road  
Bangalore 560059, India  
ksragav@hotmail.com
Raja, Nalini A.  
JBIMS Library, University of Mumbai  
(SOUL Coordinator -Maharashtra and Goa), 302, Hari-Kripa, L T Road, Borivli (W) Mumbai 400 092  
nalinir99@gmail.com

Rubinow, Sara  
School of Information & Library Science  
Pratt Institute, 144 West 14th Street  
New York, 10011  
srubinow@pratt.edu

Salah, Almila Akdag  
e-Humanities Group, Royal Netherlands Academy of Arts and Sciences - KNAW  
Joan Muyskenweg 25, 1096 CJ Amsterdam, Netherlands  
alelma@gmail.com

Sales, Rodrigo de  
Professor of Information Science  
Universidade Federal Fluminense  
Rua Praia do Flamengo, 72, Ap. 928 Flamengo CEP 22210-030,  
Rio de Janeiro, RJ Brazil  
rodrigosales@vm.uff.br

Sampalli, Tara  
Medical Informatics,  
Dalhousie University, Halifax, Canada B3H 4R2  
tara.sampalli@cdha.nshealth.ca  
tara.sampalli@dal.ca

Sasaki, Michelli  
Information Architect at UOL  
Universidade de São Paulo, Brazil  
michellichu@gmail.com

Scharnhorst, Andrea  
e-Humanities Group, Royal Netherlands Academy of Arts and Sciences  
Anna van Saksenlaan 10  
2593 HT The Hague  
andrea.scharnhorst@dans.knaw.nl

Segundo, Rosa San  
Universidad Carlos III de Madrid  
C/ Madrid 126, Getafe (Madrid) 28903, Spain  
rsan@bib.uc3m.es

Sharada, B.A.  
21, I Stage, Gangothri Layout  
Mysore 570 009, India  
sharadamallik@gmail.com

Sharatha, B.  
Indian Institute of Science  
Bangalore 560 012, India

Shepherd, Michael  
Faculty of Computer Science  
Dalhousie University  
Goldberg Computer Science Building  
6050 University Avenue, PO Box 15000  
Halifax, NS, Canada B3H 4R2  
shepherd@cs.dal.ca

Silva, Daniela Lucas da  
School of Applied Mathematics  
Getulio Vargas Foundation  
Rio de Janeiro, Brazil  
danielalucas@hotmail.com

Silva, Marco Donizete Paulino da  
School of Library and Information Science UFSCar  
Rua Martin Luther King  
1694 Boa Vista, CEP 13575-020  
São Carlos-SP, Brazil  
marco_donizete@yahoo.com.br

Smiraglia, Richard P.  
School of Information Studies University of Wisconsin – Milwaukee  
Milwaukee, USA  
smiragli@uwm.edu

Souza, Renato Rocha  
School of Applied Mathematics  
Getúlio Vargas Foundation  
Praia de Botafogo, 190  
Rio de Janeiro RJ, Brazil  
rsouza.fgv@gmail.com
Suchecki, Krzysztof  
e-Humanities Group, Royal Netherlands Academy of Arts and Sciences - KNAW  
Joan Muyskenweg 25, 1096 CJ  
Amsterdam, Netherlands  
ksucheck@gmail.com

Szostak, Rick  
University of Alberta, Edmonton, AL T6G 2H4 Canada  
rick.szostak@ualberta.ca

Szulc, Jolanta  
Institute of Librarianship and Scientific Information  
University of Silesia, Poland  
jolanta.szulc@us.edu.pl

Tanti, Marc  
University of Aix-Marseille  
Medical Intelligence Unit  
French Forces Institute of Tropical Medicine BP 46-13262 Marseille  
Cedex 07, France  
mtanti@gmx.fr

Tennis, Joseph T.  
Information School  
University of Washington, Seattle, USA  
jtennis@u.washington.edu

Togrul, V. K.  
The WB National University of Juridical Sciences  
Salt Lake City, Kolkata 700 098  
vkthoma@gmail.com; library@nujs.edu

Tilley, Carol L.  
School of Library and Information Science  
University of Illinois at Urbana-Champaign, 501 E. Daniel Street  
MC-493, Champaign  
IL 61820-6211 USA  
klabarre@illinois.edu

Tinelli, Marcele  
School of Library and Information Science UFSCar  
Rua Coronel Arthur Whitacker, 1488, São Benedito, CEP 13690 000  
Descalvado-SP, Brazil  
marceletinelli@yahoo.com.br

Tirilly, Pierre  
School of Information Studies, University of Wisconsin Milwaukee 3210 N. Maryland Avenue  
Milwaukee, WI 53201, USA  
tirilly@uwm.edu

Tognoli, Natália Bolfarini  
School of Information Science  
São Paulo State University, (UNESP)  
Marília- SP Brazil  
nataliafbtognoli@yahoo.com.br

Vinod Kumar, B.L.  
Swami Vivekananda Yoga Anusandhanam  
Samasthana, Prashant Kutir  
Bangalore, India  
vindukumarbl@gmail.com

Vogel, Michely Jabala Mamede  
Universidade de São Paulo, Brazil  
michelyvogel@usp.br

Wadekar, M.P.  
Y. M. College  
Pune 411 038, India

Wang, Mi  
Wee Kim Wee School of Communication & Information, Nanyang Technological University 31 Nanyang Link, Singapore 637718  
wangmi19830921@gmail.com

Watters, Carolyn  
Academic and Provost  
Dalhousie University, Henry Hicks Building, 6299 South Street Halifax  
NS, Canada, 902-494-2586  
cwatters@dal.ca
Will, Leonard
Willpower Information
Information Management Consultants
27 Calshot Way, Enfield, EN2 7BQ, UK
l.will@willpowerinfo.co.uk

Yun, Xin Jie
Wee Kim Wee School of Communication & Information
Nanyang Technological University
31 Nanyang Link Singapore 637718
cazydiamond@yahoo.cn

Zeng, Marcia Lei
School of Library and Information Science
Kent State University, Kent
OH 44242-0001, USA
mzeng@kent.edu

Zhang, Dong
Wee Kim Wee School of Communication & Information
Nanyang Technological University
31 Nanyang Link Singapore 637718
w090026@hotmail.com

Žumer, Maja
University of Ljubljana, Aškerčeva 2
SI-1000 Ljubljana, Slovenia
maja.zumer@ff.uni-lj.si
Author Index

Amaral, Roniberto Morato, 312
Arastoopoor, Sholeh, 345
Arboit, Aline Elis, 44
Asundi, A.Y., 166, 171
Ávila, Daniel Martinez, 206, 368
Babik, Wieslaw, 291
Barros, Thiago Henrique Bragato, 272
Bernauoi, Radia, 359
Bhanumurthy, K., 153
Bufrem, Leilah Santiago, 44
Campbell, Grant D., 195
Cerdeira, Pablo de Camargo, 320
Chada, Daniel Magalhães, 320
Chen, Hsueh-hua, 249
Chen, Shu-jiun, 249
Chudamani, K. S., 377
Codoço, Claudia Torres, 138
Coelho, Flávio Codeço, 138, 262, 320
Cordeiro, Rosa Inês de Novais, 229
Correia, Carlos Alberto, 138, 262, 320
D’Ambrosio, Amber, 236
Dahibhate, N.B., 379
Dasgupta, Abhijit, 356
Dayani, Mohammad Hosien, 378
Deokattey, Sangeeta, 153
Dhapte, Vividha, 379
Dixit, D.K., 153
Dodehei, Vera, 103, 256, 366
El Hadi, Widad Mustafa, 220
Fan-Hua, Nan, 298
Faria, Leandro Innocentini Lopes de, 312
Fattahi, Rahmatollah, 345, 371, 378
Favier, Laurence, 220
Fox, Melodie, 352
Gao, Cheng, 51
Gontijo, Miriam, 103, 256, 366
Gou-Chi, Peter, 15
Hajibayova, L., 159
Hassoun, Mohamed, 359
Henttonen, Pekka, 277
Higuchi, Suemi, 262
Hjørland, Birger, 8
Hoffmann, Wanda Aparecida Machado, 312
Iyer, Hemalata, 236
Jacob, E. K., 159
Julien, Charles-Antoine, 109, 117
Kadam, S.S., 379
Karanth, Pallavi, 90
Khoo, Christopher S.G., 83
Khosravi, Abdolrasoul, 371, 378
Kohashi, Nair Yumiko, 326, 376
Krishnamurthy, M., 145
Kulkarni, B.A., 379
Kulkarni, Umesh, 379
La Barre, Kathryn, 229, 338
Leide, John E., 109
Lima, Gercina Ângela Borém de Oliveira, 96, 124
Lorenzon, Emilena Josemari, 312
Maculan, Benildes Coura Moreira dos Santos, 124
Mahesh, Kavi, 90
Milonas, Elizabeth, 201
Ming-Shu, Yuan, 298
Mitchell, Joan S., 190
Moraes, João Batista Ernesto de, 242, 272
Mugade, V.S., 379
Murguia, Eduardo Ismael, 183
Nagarathna, H.C., 377
Neelameghan, A., 176, 212
Nikam, Khaiser, 306
Nizam, Naureen, 69
Ohly, Peter, 15
Oliveira, Ely Tannuri de, 31, 44
Orrico, Evelyn, 103, 256, 366
Padmavathi, T., 145
Parikkh, Mehr, 371, 378
Pattuelli, Maria Cristina, 75
Raghavan, K.S., 176, 212
Raja, Nalini A., 333
Rubinow, Sara, 75
Sales, Rodrigo de, 183
Sampalli, Tara, 131
Sasaki, Michelli, 376
Scharnhorst, Andrée, 51
Segundo, Rosa San, 206
Sharada, B.A., 63
<table>
<thead>
<tr>
<th>Name</th>
<th>Page(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharatha, B.</td>
<td>377</td>
</tr>
<tr>
<td>Shepherd, Michael</td>
<td>131</td>
</tr>
<tr>
<td>Silva, Daniela Lucas da</td>
<td>262</td>
</tr>
<tr>
<td>Silva, Marco Donizete Paulino da</td>
<td>312</td>
</tr>
<tr>
<td>Smiraglia, Richard P.</td>
<td>1, 51</td>
</tr>
<tr>
<td>Souza, Renato Rocha</td>
<td>138, 262, 320</td>
</tr>
<tr>
<td>Suchecki, Krzysztof</td>
<td>51</td>
</tr>
<tr>
<td>Szostak, Rick</td>
<td>24</td>
</tr>
<tr>
<td>Szule, Jolanta</td>
<td>369</td>
</tr>
<tr>
<td>Tanti, Marc</td>
<td>372</td>
</tr>
<tr>
<td>Tennis, Joseph T.</td>
<td>39, 58</td>
</tr>
<tr>
<td>Thomas, V. K.</td>
<td>365</td>
</tr>
<tr>
<td>Tilley, Carol L.</td>
<td>338</td>
</tr>
<tr>
<td>Tinelli, Marcele</td>
<td>312</td>
</tr>
<tr>
<td>Tirilly, Pierre</td>
<td>109, 117</td>
</tr>
<tr>
<td>Tognoli, Natália Bolfarini</td>
<td>267</td>
</tr>
<tr>
<td>Vinod Kumar, B.L.</td>
<td>306</td>
</tr>
<tr>
<td>Vogel, Michely Jabala Mamede</td>
<td>376</td>
</tr>
<tr>
<td>Wadekar, M.P.</td>
<td>379</td>
</tr>
<tr>
<td>Wang, Mi</td>
<td>83</td>
</tr>
<tr>
<td>Watters, Carolyn</td>
<td>69</td>
</tr>
<tr>
<td>Will, Leonard</td>
<td>284</td>
</tr>
<tr>
<td>Yun, Xin Jie</td>
<td>83</td>
</tr>
<tr>
<td>Zeng, Marcia Lei</td>
<td>190, 249</td>
</tr>
<tr>
<td>Zhang, Dong</td>
<td>83</td>
</tr>
<tr>
<td>Žumer, Maja</td>
<td>190</td>
</tr>
</tbody>
</table>
SUBJECT INDEX

Aboutness
  *irt* Images, 345, 346, 349, 351, 353
Access points, 229, 234, 236
Agricultural Information Systems, 359
Agriculture Schedule
  *irt* Colon Classification, 171, 174
AGROVOC
  (FAO Thesaurus), 359, 365
Alignment
  *irt* Controlled Vocabularies, 249
Alternative Subject Access Fields, 333
Alzheimer’s Disease Information
  *irt* Hyperlinking, 195-197, 199, 201, 202
Archetypes, 236, 237, 243
Archives, Classification, 267, 272, 275, 279
Authors
  *irt* Main Thematic Categories, 44
Automatic Indexing, 326, 329
Basic Concepts Classification, 24-30
Basic Level Categories, 159
BCC
  *see* Basic Concepts Classification
Bilingual Thesaurus, 306, 309
Boundary Object
  *irt* Ontology, 131
Cartographic Materials, 291, 294, 296-299
CATIA
  *see* Computer Aided Three-dimensional Interactive Application
CC
  *see* Colon Classification
CDA
  *see* Critical Discourse Analysis
CNPq (Brazil), 183, 184, 187, 188, 189, 190
Chinese Art, 249
Citation analysis
  *irt* Indexing research, 39
Citation Frontiers
  *irt* Indexing Theory, 39
Classification, 377
  Systems, 51
Cloud Computing, 377
Cognitive Load Theory, 371, 378
Collaborative tagging
  *irt* Gender and Social Tagging, 352
Colon Classification, 58, 61, 63-70
Computer Aided Three-dimensional Interactive Application, 368, 369
Concept, 212, 213, 216, 219
  *irt* Knowledge organization, 176, 182
Conceptual classification
  *irt* Archival Science, 272
  modeling, 96
  structures
  *irt* Digital Environment, 206, 249
Controlled Vocabularies, 312
Controlled Vocabulary, 117, 312
  *irt* of E-Government - VCGE, 256, 257
Co-word Analysis, 153, 158, 159
Critical Discourse Analysis, 256, 260
Cultural Frames, 212
Culture
  *irt* Knowledge Organization, 212, 213, 215-217, 218
Data Model
  *irt* Information Retrieval Thesaurus, 284
Data Visualization, 320
Digital Libraries
  *irt* Theses and Dissertations, 96, 124
Digitized Content, 333
Dimensions, 1, 7
Discourse Analysis, 256, 260, 264
Documentary universalism, 220, 228
Domain Analysis
  irt Knowledge Organization, 1, 7, 9, 31, 34, 298, 310
  Specific Categories, 171
Dublin Core, 368
Ecological change
  irt Faceted Classification, 58, 61, 64, 65
Electronic Documents
  irt Naming, 372, 373
Elitism analysis of Price, 39
Epistemological basis
  irt Common Categories, 166
classification, 31
Epistemology of knowledge organization, 31, 40
Ervin Panofsky’s Theory, 345
Facet Analysis, 153, 229, 230, 232, 234, 237
Faceted classification, 58, 291
taxonomy, 124
Facets of Law, 365
Fiction, 242, 250
Films and Facets
Fishery industry, 298, 301, 304-309
Food Science and Technology, 145, 148
FRBR
  190, 193, 197, 198, 338, 341, 344, 346-348
Full-Text Access, 338
Functional Requirements for Bibliographic Records,
  see FRBR
Gender and Social Tagging, 352
Generative Trajectory of Meaning, 242, 246, 248
Google Suggest, 371
Google's Suggested Keywords, 378
Greenstone Digital Library, 306, 309
  irt machine-readable Sanskrit-English thesaurus, 306, 309, 311, 313
GSDL
  see Greenstone Digital Library
Hypertext Map Prototype, 96, 99, 102-106
Hypertexts, 96
ICM
  see Idealized Cognitive Models
Idealized Cognitive Models, 236, 237, 241, 243
Image organization, 236, 237
Index Pages, 333
Indic Cultures, 176, 182
Industrial Technology, 298
Inference Engine RACER, 369
Information architecture, 8
mining, 320
organization, 8, 10, 12-15, 374
resources, 83, 89
retrieval 103, 109, 110, 374
filter, 103, 106, 109
visualization, 326
InfoSIC (Competitive Intelligence System), 312-318
Intelligent Information System, 312
International Institute of Bibliography, 220
IO
  see Information organization
ISO 25964-1: 2011 – Thesauri for information retrieval, 284
Kannada-English Version, 63-68
  see Colon Classification
KAT
  see Knowledge Area Table
Keyword Suggests (Google), 371
Knowledge and Power Apparatus, 183
area Table, 183, 186-188
classification Scheme, 298
organization
  irt Records and Archives, 1-16,
  19, 20, 23-25, 31, 34, 37,
  40-42, 69, 159, 167, 176, 242,
  249, 272, 275, 278-280, 368
tools, 176, 179, 218, 219
Management, 277, 280, 284, 365
journal, 31
scheme, 90
tools, 176, 179, 223
representation, 267, 366
KO
  see Knowledge Organization
KOS Relationships, 190
KOT
  see Knowledge organization tools
LCSH
  see Library of Congress Subject Headings
Lexical change
  irt Faceted Classification, 58, 62, 64
Library of Congress Subject Headings, 109, 110, 112, 113, 115-121
browsing
  irt Information retrieval, 109
Linked
  Dataset, 75
  Open Data, 75, 84
LOD
  see Linked Open Data
Main Thematic Categories, 44
Mathematical
  Epidemiology, 138
  Modeling, 138, 143
Metaphor, 103, 107
MHTX
  see Hypertext Map Prototype
Micro-domain, 153
Moving image, 229, 234, 237
repositories, 229, 234, 235
Narrative texts of fiction, 242
News Portal, 374
Ofness
  irt Images, 345, 346, 349, 351, 353
OI
  see Organization of information
Ontological Representation
  irt Knowledge, 145
Ontology, 131, 134, 138, 139, 141,
  143-146, 153, 160, 161, 369
construction, 256
dictionary, 369
Organization of collections, 379
information, 8, 9, 11-13
Otlet, Paul, 220, 223, 226, 228-232
Panofsky’s Theory, 345
Paraconsistent Logic
  irt Automatic Indexing, 326
PEx
  see Projection Explorer
Postal Stamps, 379
Projection Explorer, 326, 331-335
Query Expansion, 371, 378
RACER
  see Renamed ABox and Concept Expression Reasoner
Relational
Indexing, 195
operator of Farradane, 195
Relevance performance, 378
Renamed A Box and Concept Expression Reasoner, 369, 370
Search engines, 201, 202, 204, 206, 207
Self Organizing Knowledge Organization Systems, 15
Semantic and Knowledge Organization Systems, 262
approaches, 15, 23
content, 96
content management
  Digital libraries, 96, 99
Interconnection, 206, 210, 212
interoperability, 131
portal project
  Semantic and Knowledge Organization Systems, 262
relationships, 212
roles, 90, 94
Web Technology, 365
SNOmED CT®
  Ontology, 131-134, 137
Social
classification, 69
identity Categories, 366
Social Tagging, 352, 356, 357
Space, 166, 169, 170, 172
Subject
analysis, 242
hierarchy, 117
indexing
  Images, 345, 353
Organization, 83
Superlinks
  Hyper linking structure, 90
Technological knowledge
classification scheme, 298, 301, 304-306, 308, 309
Terms of Non-Latin Origin, 306
Text and image documents, 220
The Least Effort Principle, 378
Thesaural relationships, 201, 204, 206
Time, 166, 169-172
effect, 58
  Knowledge organization tools, 212-216, 218
Tools for Conceptualization, 153
UDC
  see Universal Decimal Classification
Uncontrolled terms, 334, 336
Universal Decimal Classification, 52, 62
  Ontogeny, 51, 52, 54, 55, 57-62
Universalist philosophy, 220
Universality, 220, 227
Universes, 1
User Tasks
  FRBR, 338
Web 3.0, 15, 23, 26, 69, 71, 72, 74, 90, 93-97
Search engines, 371
Term Relationships, 201
Yogic Sciences, 306
Twelfth International Conference of the International Society for Knowledge Organization (ISKO) was held at the University of Mysore, Mysore, India from 6th to 9th August 2012.

Organizers
ISKO (India Chapter)
Sarada Ranganathan Endowment for Library Science (SRELS)
University of Mysore

Co-Sponsors
Defence Research & Development Organization (DRDO), Government of India
PES Institute of Technology, Bangalore
Ranganathan Centre for Information Studies, Bangalore
Society for Information Science (Bangalore Chapter)
University of Madras, Department of Library & Information Science, Chennai

Organizing Committee
Chair
V.G. Talawar, Vice-Chancellor, University of Mysore

Programme Chair
A. Neelameghan (SRELS, Bangalore)

Convener / Organizing Secretary
K. S. Raghavan (Indian Statistical Institute, Bangalore)

Rapporteur General
Kathryn La Barre (University of Illinois- Urbana Champaign)

Members
A. Amudhavalli (University of Madras, Chennai)
Kavi Mahesh (PES Institute of Technology, Bangalore)
P.S. Naik (University of Mysore, Mysore)
S. Parthasarathy (Ranganathan Centre for Information Studies, Chennai)
K.N. Prasad (SRELS, Bangalore)
I.K. Ravichandra Rao (SRELS, Bangalore)

Website & Technical Support
H.M. Rajashekara (Indian Statistical Institute, Bangalore)
International Scientific Committee

Wieslaw Babik, University in Cracow, Poland
Clare Beaghtol, University of Toronto, Canada
Ingetraut Dahlberg, Germany
B.S. Dayasagar, Indian Statistical Institute (Bangalore), India
Widad Mustafa El Hadi, UFR IDIST - Labo. GERIICO, University of Lille3, France
Ibekwe-Sanjuan Fidelia, University of Lyon3, France
Birger R. Hjorland, Royal School of Library and Information Science, Copenhagen
Claudio Gnoli, University of Pavia, Italy
Rebecca Green, OCLC, USA
Jane Greenberg, University of North Carolina at Chapel Hill, USA
Jose Augusto Guimaraes, Universidade Estadual Paulista, Sao Paulo
Hemalatha Iyer, University at Albany, State University of New York, USA
Christopher Khoo, Nanyang Technological University, Singapore
Barbara H. Kwasnik, Syracuse University
Kathryn La Barre, University of Illinois, Urbana Champaign, USA
Maria Lopez, University of Granada, Spain
Kavi Mahesh, PES Institute of Technology, Bangalore
Fulvio Mazzocchi, National Research Council, Italy
Dorothy McGarry, University of California at Los Angeles, USA
Ia Mcllwaine, University College London, UK
Kavi Narayana Murthy, University of Hyderabad, India
A. Neelameghan, Sarada Ranganathan Endowment for Library Science, India
Peter Ohly, ISKO, Germany
Hope Olson, University of Wisconsin Milwaukee, USA
Vivien Petras, Humboldt-University, Berlin, Germany
K. S. Raghavan, Indian Statistical Institute (Bangalore), India
I. K. Ravichandra Rao, Indian Statistical Institute (Bangalore), India
M.P. Satija, Gurunanak Dev University, Amritsar, India
Michael Shepherd, Dalhousie University, Canada
Aida Slavic, UDC Consortium, London, UK
Richard P. Smiraglia, University of Wisconsin, Milwaukee, USA
Joseph Tennis, University of Washington, USA
Carolyn Watters, Dalhousie University, Canada
Irene Wormell, Denmark
Marcia Zeng, Kent State University, USA