Culture and Identity in Knowledge Organization
The volume contains:
# Table of Contents

## INTRODUCTION

### KEYNOTE ADDRESS

**Jonathan Furner.** Interrogating “Identity”: A Philosophical Approach to an Enduring Issue in Knowledge Organization.

## MODELS AND METHODS IN KNOWLEDGE ORGANIZATION

**Louise F. Spiteri.** Causality and Conceptual Coherence in Assessments of Similarity.

**Rebecca Green.** Making Visible Hidden Relationships in the Dewey Decimal Classification: How Relative Index Terms Relate to DDC Classes.

**John DiMarco.** Examining Bloom’s Taxonomy and Peschl’s Modes of Knowing for Classification of Learning Objects on the PBS.org/teachersource Website.

**Melanie Feinberg.** Classificationist as Author: The Case of the Prelinger Library.

**Fulvio Mazzocchi and Mela Bosch.** Hermeneutic Approaches in Knowledge Organization: An Analysis of Their Possible Value.

**Yves Marcoux and Élias Rizkallah.** Knowledge Organization in the Light of Intertextual Semantics: A Natural-Language Analysis of Controlled Vocabularies.

**Vanda Broughton.** Language Related Problems in the Construction of Faceted Terminologies and their Automatic Management.

**I. C. McIlwaine and Nancy J. Williamson.** Medicine and the UDC: The Process of Restructuring.

**James M. Turner.** Cultural Markers and Localising the MIC Site.

**João Alberto de Oliveira Lima, Monica Palmirani and Fabio Vitali.** A Time-aware Ontology for Legal Resources.

## MULTILINGUAL AND MULTICULTURAL ENVIRONMENTS


**Elaine Ménard.** Indexing and Retrieving Images in a Multilingual World.

**Maria Odaisa Espinheiro de Oliveira.** Knowledge Representation Focusing Amazonian Culture.

**Ágnes Hajdu Barát.** Knowledge Organization in the Cross-cultural and Multicultural Society.

**Joan S. Mitchell, Ingebjorg Rype and Magdalena Svanberg.** Mixed Translation Models for the Dewey Decimal Classification (DDC) System.
<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNOWLEDGE ORGANIZATION FOR LIBRARIES, ARCHIVES AND MUSEUMS</td>
<td></td>
</tr>
<tr>
<td>Kathryn La Barre. Discovery and Access Systems for Websites and Cultural Heritage Sites: Reconsidering the Practical Application of Facets.</td>
<td>105</td>
</tr>
<tr>
<td>Mats Dahlström and Joacim Hanson. On the Relation Between Qualitative Digitization and Library Institutional Identity.</td>
<td>111</td>
</tr>
<tr>
<td>Amelia Abreu. “Every Bit Informs Another”: Framework Analysis for Descriptive Practice and Linked Information.</td>
<td>118</td>
</tr>
<tr>
<td>Jenn Riley. Moving from a Locally-developed Data Model to a Standard Conceptual Model.</td>
<td>124</td>
</tr>
<tr>
<td>Jan Pisanski and Maja Žumer. How Do Non-librarians See the Bibliographic Universe?</td>
<td>131</td>
</tr>
<tr>
<td>KNOWLEDGE ORGANIZATION FOR INFORMATION MANAGEMENT AND RETRIEVAL</td>
<td></td>
</tr>
<tr>
<td>Xu Chen. The Influence of Existing Consistency Measures on the Relationship Between Indexing Consistency and Exhaustivity.</td>
<td>144</td>
</tr>
<tr>
<td>Michael Buckland and Ryan Shaw. 4W Vocabulary Mapping Across Diverse Reference Genres.</td>
<td>151</td>
</tr>
<tr>
<td>Abdus Sattar Chaudhry and Christopher S. G. Khoo. A Survey of the Top-level Categories in the Structure of Corporate Websites.</td>
<td>157</td>
</tr>
<tr>
<td>Nicolas L. George, Elin K. Jacob, Liqiang Guo, Lala Hajibayova and M. Yasser Chuttur. A Case Study of Tagging Patterns in del.icio.us.</td>
<td>164</td>
</tr>
<tr>
<td>Kwan Yi and Lois Mai Chan. A Visualization Software Tool for Library of Congress Subject Headings.</td>
<td>170</td>
</tr>
<tr>
<td>Gercina Ângela Borém Oliveira Lima. Hypertext Model — HTXM: A Model for Hypertext Organization of Documents.</td>
<td>177</td>
</tr>
<tr>
<td>Verónica Vargas and Catalina Naumis. Water-related Language Analysis: The Need for a Thesaurus of Mexican Terminology.</td>
<td>190</td>
</tr>
<tr>
<td>Amanda Hill. What’s in a Name?: Prototyping a Name Authority Service for UK Repositories.</td>
<td>196</td>
</tr>
<tr>
<td>Rick Szostak and Claudio Gnoli. Classifying by Phenomena, Theories and Methods: Examples with Focused Social Science Theories.</td>
<td>203</td>
</tr>
</tbody>
</table>
EPISTEMOLOGICAL FOUNDATIONS OF KNOWLEDGE ORGANIZATION

H. Peter Ohly. Knowledge Organization Pro and Retrospective. 210

Judith Simon. Knowledge and Trust in Epistemology and Social Software/Knowledge Technologies. 216

D. Grant Campbell. Derrida, Logocentrism, and the Concept of Warrant on the Semantic Web. 222

Jian Qin. Controlled Semantics Versus Social Semantics: An Epistemological Analysis. 229

Hope A. Olson. Wind and Rain and Dark of Night: Classification in Scientific Discourse Communities. 235


Richard P. Smiraglia. Noesis: Perception and Every Day Classification. 249

Birger Hjørland. Deliberate Bias in Knowledge Organization? 256

Joseph T. Tennis and Elin K. Jacob. Toward a Theory of Structure in Information Organization Frameworks. 262

Jack Andersen. Knowledge Organization as a Cultural Form: From Knowledge Organization to Knowledge Design. 269

Hur-Li Lee. Origins of the Main Classes in the First Chinese Bibliographic Classification. 275

NON-TEXTUAL MATERIALS

Abby Goodrum, Ellen Hibbard, Deborah Fels and Kathryn Woodcock. The Creation of Keysigns American Sign Language Metadata. 282


DISCOURSE COMMUNITIES AND KNOWLEDGE ORGANIZATION

Aaron Loehrlein. The Benefits of Participating in a Form of Life: Interpretations of Complex Concepts among Experts and Novices in Records Management. 295

Widad Mustafa El Hadi. Discourse Community Analysis: Sense Construction Versus Non-Sense Construction. 302

**Users and Social Context**

*Edward M. Corrado* and *Heather Lea Moulaison*. Social Tagging and Communities of Practice: Two Case Studies. 313

*Margaret E. I. Kipp*. Searching with Tags: Do Tags Help Users Find Things? 320

*Lynne C. Howarth*. Creating Pathways to Memory: Enhancing Life Histories through Category Clusters. 326

*Christopher S. G. Khoo* and *Shiyan Ou*. Machine Versus Human Clustering of Concepts Across Documents. 333

*Maria J. López-Huertas*. Cultural Impact on Knowledge Representation and Organization in a Subject Domain. 340

*Inge Alberts* and *Suzanne Bertrand-Gastaldy*. A Pragmatic Perspective of E-mail Management Practices in Two Canadian Public Administrations. 347

*June Abbas*. Daddy, How do I Find a Book on Purple Frogs?: Representation Issues for Children and Youth. 354

*José Augusto Chaves Guimarães, Juan Carlos Fernández-Molina, Fabio Assis Pinho* and *Suellen Oliveira Milani*. Ethics in the Knowledge Organization Environment: An Overview of Values and Problems in the LIS Literature. 361

**Systems, Tools and Evaluation**

*Ismaïl Timimi* and *Stéphane Chaudiron*. Information Filtering as a Knowledge Organization Process: Techniques and Evaluation. 367


**List of Contributors** 381

**Subject Index** 387
Introduction

This 10th International Conference on Knowledge Organization research has as its theme, *Culture and Identity in Knowledge Organization*. The theme was chosen because of the site of the conference. Culture and identity are a powerful component of life in the Province of Québec. Montréal is synonymous with the concepts of culture and identity. In this city we find the Bibliothèque et Archives nationales du Québec, museums and archives attesting to a history distinct from, though in constant interaction with, the Anglo-dominated conception of North America. Beyond the city we see the power of culture and identity at work in Québec. Nunavik, in the north of this province is working toward Inuit self-governance. Odanak, home of the Abenakis aboriginal people hosts one of Québec’s largest Aboriginal Museums. Thus, as we can see, cultural heritage and the identities to which it attests are part of the fabric of this place. This ethos, in situ here in Montréal, shaped our call for papers.

The papers presented in this volume, responding to our call, span nine themes: models and methods in knowledge organization; multilingual and multicultural environments; knowledge organization for libraries, archives and museums; knowledge organization for information management and retrieval; epistemological foundations of knowledge organization; non-textual materials; discourse communities and knowledge organization; users and social context; systems, tools and evaluation. The theme of the conference is present throughout the papers—many addressing multicultural issues, and issues of culture and identity in scholarly disciplines.

And it is in that vein that I want to comment that the state of research in Knowledge Organization is also about culture and identity. In helping to put this volume together I have grown increasingly aware that we occupy a particular space in the contemporary landscape. There are many communities working on problems of organizing the world’s knowledge, each with their own language; each with its own priorities, values, and ways of working. We within the province of Knowledge Organization are but one culture at work in this landscape of research. We border lands of ontology engineering, enterprise information management, social theory, and personal information management. Some of us even visit these realms to do our work. And while we may have left lands of research in descriptive bibliography, we are fast moving into the territory of folksonomy. Where we go in this *terra semantica* is up to us, but regardless of where we tread, we know we will not tread alone. And how we decide where to venture will be based on our personal and communal conceptions of the identity and culture of Knowledge Organization.

I’d like to conclude by offering thanks to those who helped make this publication possible. Thanks go first to the scholars that contributed excellent papers to this conference. Thanks also go to the reviewers who took time and effort to suggest improvements to the submis-
sions. My thanks join a chorus of thanks from the submitting authors. Finally, sincerest thanks go to Dr. Clément Arsenault of the Université de Montréal. He deftly coordinated the production of these proceedings, set the format, coordinated the indexing with the help of his colleague Michèle Hudon, and sent it all to Ergon. Thanks to all, and I wish you a good conference, as we contemplate the role of culture and identity in Knowledge Organization.

Joseph T. Tennis
Seattle, April 30, 2008
Empirical evaluation of knowledge organization (KO) systems, and of the tools and techniques that are used to build systems, is a key component of the system design process: our success in building better systems depends at least partly on our ability to measure the goodness of current systems, and to recognize the factors that affect system performance. The basic evaluative question might be expressed quite simply: How good are the representations or models — models of the world, of our knowledge of the world, and/or of expressions of our knowledge of the world — that are produced by our usage of particular KO methods? The straightforwardness of this question is offset by a preliminary need to address metaphysical issues of various kinds, consideration of which can lead us into a quagmire of methodological, epistemological, and ethical problems. What, in this context, is “goodness”? What is the fundamental nature of the kinds of things to be represented? What are the conditions that must be satisfied for a single individual thing to retain its identity over time, and for two individual things to be instances of “the same” kind of thing? Where are the boundaries to be drawn between one thing (or kind of thing) and another? Where does one thing (or kind of thing) stop and another start? How can we come to know the answers to questions about identity, and how we can know when we know? How have we answered questions about identity in different ways at different times and in different places? How ought we to answer questions about identity, and what justifications can we provide in support of our normative claims? As is indicated by the conference organizers’ choice of theme for ISKO 2008, designers and evaluators of KO schemes contend on an ongoing basis with issues relating to identity, and a philosophically-informed engagement with such issues is an essential preliminary to understanding evaluation criteria for KO activity. In this talk, the utility for KO of philosophical theories of identity is examined, and motivation is provided for the additional use of such a philosophical framework in evaluating the extent to which KO schemes successfully reflect the cultural identities of their users.
Causality and Conceptual Coherence in Assessments of Similarity

Abstract
Conceptual coherence, which refers to concepts whose contents make sense to the perceiver, has been associated traditionally with the notion of similarity, that is, objects, events, or entities form a concept because they are similar to one another. An examination of traditional similarity-based concept theories suggests that they do not provide an adequate account for conceptual coherence. Library and Information Science needs to explore knowledge-based approaches to concept formation, which suggest that one’s knowledge of a concept includes not just a representation of its features, but also an explicit representation of the causal mechanisms that people believe link those features to form a coherent whole.

Introduction
“Concepts are the glue that holds our mental world together […]. Our concepts embody much of our knowledge of the world, telling us what things there are and what properties they have” (Murphy, 2002, p. 1). The standard psychological usage of concept is that of a mental representation individuated or defined by its contents or features (Laurence & Margolis, 1999). Concepts are tied closely to categories: Categorization involves characterizing something by means of concepts so, for example, my concept of dog allows me to pick out a category of entities that I would call dogs. Conceptual coherence refers to concepts whose contents “seem to hang together, a grouping of objects that makes sense to the perceiver” (Murphy & Medin, 1999, p. 427). Conceptual coherence has been associated traditionally with the notion of similarity, that is, objects, events, or entities form a concept because they are similar to one another. Objects fall into natural clusters of similar kinds (e.g., dogs) that are, at the same time, dissimilar to other kinds (e.g., cats).

The notion of similarity, or likeness, underlies many approaches used in Library and Information Science (LIS) in the design of bibliographic classification systems: The reliance upon similarity assumes a shared or common understanding of the attributes or features that give a concept its identity. Does similarity explain, however why a concept was formed or why it makes sense to the perceiver? Will the same concept have the same degree of coherence amongst different people, even within the same domain? Prior theoretical knowledge may contribute to representations of categories; for example, people not only know that birds have wings and that they can fly and build nests in trees, but also that birds build nests in trees because they can fly, and fly because they have wings (Rehder, 2003). In comparison, however, with the development of models that account for the effects of similarity and empirical observations, there has been relatively little development of formal models to account for the effects of such prior knowledge on the formation of concepts.

This paper will examine how conceptual coherence is defined and explored in existing concept theories. It will be argued that traditional similarity-based theories do not provide an adequate account for conceptual coherence, and that LIS needs to explore newer, knowledge-based approaches to concept formation, which suggest that one’s knowledge of many concepts includes not just a representation of a concept’s features, but also an explicit representation of the causal mechanisms that people believe link those features to form a coherent whole.
In the cognitive sciences, similarity is thought to play an essential role in how people acquire and categorize information: Once knowledge is acquired, similarity serves as an organizing principle by which individuals classify objects, form concepts, and make generalizations (Tversky, 1977). Similarity is the foundation of inductive thinking, since categories whose members share similar properties have stronger inductive power than categories whose members are less similar (Heit, 1997; Murphy, 2002).

The concept of similarity, normally referred to as likeness in the LIS literature, is often stated as being an important underlying principle of bibliographic classification. Shera (1965) posits four basic assumptions of bibliographic classification: universal order of knowledge; hierarchical (genus-species) division; differentiation; and permanence. Maltby defines classification as “not only the grouping of things which resemble one another and the separation of those which do not, but the arrangement within each group of its components according to their degree of resemblance” (Maltby, 1975, p. 16). Broadfield (1946), however, suggests that likeness indicates merely a relationship between things; it is not a characteristic of things.

**Similarity-based theories of concept formation**

The Classical Theory posits that concepts contain necessary and sufficient conditions by which they can be defined. Categorization is a process of checking to see if the features that are part of a concept are satisfied by the item being categorized: Entities that are considered similar are members of the same category by virtue of the fact that they share the same properties. All members of the same category are equally similar to each other because they possess the same properties; similarity is thus symmetrical, because what is true for one entity in the category is true also for another.

The Prototype Theory argues that all concepts show gradient degrees of membership; for example, a sparrow is a better example of *bird* than is an emu, because a sparrow is associated more readily with the features that one attributes to birds (Rosch, 1999). Items can be considered extremely typical, moderately typical, atypical, and borderline concept members; typicality is thus a graded phenomenon. The judged similarity of any two items is measured by comparing the sets of shared and distinctive features that are associated with them. Prototype Theory is sensitive also to context; for example, dogs or cats might be cited as prototypical pet animals, while elephants and lions prototypical circus animals.

The Classical and Prototype theories both focus upon unitary descriptions that capture the central tendency of any given concept; the difference lies in the acceptance or rejection of a set of necessary and sufficient features to create that description. The Exemplar Theory suggests that people do not have a unitary definition of the concept *dog*, for example, nor is this concept composed of a list of features that is found to varying degrees amongst dogs. Rather, one’s concept of *dog* is composed only of the set of dogs that one has actually encountered and remembered (Smith & Medin, 1999). This means that my definition of *dog* would be based not on a unitary description that would necessarily apply to a majority of dogs, but to my exemplar of *dog*.

In his seminal study of similarity, Goodman (1972) concluded that saying that two things are similar does not say very much about them, since any two things can be regarded as similar or dissimilar, depending on which features one selects for the purposes of comparison. A coherent concept is one that makes sense to the perceiver: The reason that *bird* is a useful concept is that birds are relatively similar to each other; most birds
have wings, lay eggs, and fly, for instance. Items are said to belong to the same concept if they share common properties; the problem is that we may often see things as being similar because they belong to the same concept, rather than categorize them because of their similarity. Any two entities can be deemed similar or dissimilar depending on how many features one uses, and the relevance or salience that one attributes to these features (Murphy & Medin, 1999, p. 428). How well can I define what constitutes my coherent concept of dog? More importantly, is my understanding of the essence of dogness the same as other people’s? My definition of a dog may rely upon a combination of physical attributes (e.g., the ability to bark) and certain behavioural attributes (e.g., fetching, herding, or retrieving). If I place a higher value on herding behaviour, I am likely to associate this attribute more closely with barking, than a person who places a higher value on fetching behaviour, since barking is often an important component of herding. Similarity-based theories do not explain sufficiently how our underlying knowledge or understanding of the essence of a concept affects which properties we choose as well as which we combine in causal relationships to affect our understanding of a concept’s coherence.

Knowledge-based models of concept formation

Theory-Theory posits that the process of learning about most concepts involves noticing how often properties or features occur and co-occur; for example, blackness and roundness are both frequently-occurring features of tyres, yet roundness seems to be more central to tyres since it is so closely linked to the function of tyres (Keil, 2003). Concepts are learned as part of our overall understanding of the world around us; they are influenced by what we already know, but may serve also to affect our existing knowledge (Murphy, 2002; Rehder, 2003). So, for example, recent experiments in the creation of self-replicating robots could cause us to question our current understanding of the biological function of reproduction. Theory-Theory believes that we do not rely only on simple observation or feature matching to form concepts: We make inferences based upon our prior knowledge and experience and can add information that we do not observe in the item itself. The essences of concepts are not simply assumed to be defining features, but also the causal reason behind the manifestation of surface features (Prinz, 2002).

Causal-Mode Theory posits that people’s knowledge of many concepts includes not just a representation of a concept’s features, but also an explicit representation of the causal mechanisms that people believe link those features. People use causal models to determine a new object’s category membership (Rehder, 2003). Causal-Mode Theory measures the importance of individual features on concept membership based upon people’s domain theories; for example, straight bananas are rated as better members of the category bananas than straight boomerangs are of the category boomerangs, a result people attribute to the default feature curved occupying a more theoretically-central position in the conceptual representation of boomerang as compared to banana. Causal-Mode Theory posits also that particular combinations of features affect people’s decision as to what makes for a coherent concept (Rehder, 2003).

Prinz (2002) argues that with its reliance upon defining essences, the knowledge approach does not provide a sufficient explanation for conceptual structure: If I cannot identify the essence of a dog, then how can I have a coherent concept of a dog? Saying that what makes a dog is the essence of being a dog is circular at best. People’s theories may be incorrect or could change over time. The knowledge approach does not suggest that my friend and I possess the same understanding of dog, but that we are capable of
forming our own coherent concepts, which reverts to the definition of conceptual coherence, namely, that a concept make sense to the person who forms it. It is possible that my interpretation of doghood may differ from other people’s in details, but does this mean that is impossible to achieve a degree of consensus over what constitutes a dog? Rather than insist upon a unitary definition of doghood, the knowledge approach acknowledges the existence of degrees of doghood that can be agreed upon, especially within a specific domain, and the fact that these areas of consensus may vary across different domains. We must be willing to accept a degree of uncertainty and some fuzzy boundaries in the design of concepts, and that we may still find enough areas of commonalities to make concepts coherent across a domain. The concept of marriage, for example, has recently undergone changes in both societal and legal definitions in Canada, such that it no longer necessarily involves the civil union of a man and a woman. It could be argued that the knowledge approach reflects the normal progression of concepts over time that reflects changes in societal and cultural norms.

The knowledge approach accepts that one’s essence of a concept may be wrong if it is based on erroneous theories of knowledge. Once again, however, the knowledge approach may simply reflect the realities of life. It is possible for one person to accept the tenets of creationism, even if scientific evidence suggests that this theory may be unfounded. It is questionable whether similarity can prevent the formation of “wrong” concepts; a person can be provided with all the correct attributes of a concept and still choose to define the concept incorrectly. The question of consensus, however, may act as a mitigating factor; although not all the members of a domain may agree on the essence of a concept, to the point where some members’ interpretation of this concept may be perceived as “wrong,” the knowledge approach suggests that it is still possible to establish a baseline level of consensus common to the majority of the members.

Conclusion: The knowledge approach and bibliographic classification

The knowledge approach to conceptual coherence parallels well recent discussions within LIS about the structure of bibliographic classification systems. Hjørland and Albrechtsen (1999) and Beghtol (2003) argue that classification research must be situated within specific contexts and the domains in which the classification systems are designed to function. Classification is based more upon interpretation and judgment than upon logic, and its ultimate purpose is to suggest a view of the world that makes sense, or is coherent, to its users (Mai, 2004). The socio-cognitive view of domain analysis emphasizes how domains structure culturally-produced signs and symbols and how its members mediate their cognitive processes into coherent concepts that reflect shared cultural, historical, and social meanings (Hjørland, 2004). The knowledge approach’s emphasis upon consensus has parallels also within LIS. Consensus in classification can be traced to Henry Evelyn Bliss (1939), who believed that classification systems should reflect how members of the scientific and educational communities structure knowledge in their respective domains.

The dependence of many LIS bibliographic classification systems upon similarity-based measures of conceptual coherence may result in systems that impose a unitary definition of coherence on any given concept. There is a need within LIS to examine the impact of knowledge and causality upon people’s construction of concepts and to see whether it is possible to achieve a consensus of coherence for these concepts within given domains. The Causal-Model theory is the most developed working model of the knowledge approach in its formal account of how causal knowledge influences the importance
of features and specific configurations of features in judgments of concept membership. This model’s ability to provide a precise, quantitative account of both the differences in feature weights and the importance of feature configurations induced by people’s knowledge thus makes it an attractive candidate for integrating the knowledge approach into the construction of bibliographic classification systems.

References


Abstract
Relative Index (RI) terms in the Dewey Decimal Classification (DDC) system correspond to concepts that either approximate the whole of the class they index or that are in standing room there. DDC conventions and shallow natural language processing are used to determine automatically whether specific RI terms approximate the whole of or are in standing room in the classes they index. Approximately three-quarters of all RI terms are processed by the techniques described.

Introduction
The schedules of the Dewey Decimal Classification (DDC) system, ranging over notation 001–999, include over 26,000 subject classes, while the accompanying tables contribute more than 9,000 additional classes. Of a necessity, the scheme offers several types of tools for locating relevant classes, including hierarchical organization of the classes, class-elsewhere notes, see and see-also references, and the Relative Index, (which includes over 46,000 index terms for schedule entries, over 16,000 index terms for table entries, and over 10,000 index terms for built numbers). In addition to acting as a locating device, the Relative Index indicates which of multiple class numbers for a topic to use for interdisciplinary works and, to a limited extent, comprehensive works. What it does not generally do at present is indicate the relationship between the topic named by the index term and the DDC class that it indexes.

DDC classes are often gathering points for a set of topics. Some of these topics are said to “approximate the whole” of the class. While the scope of these topics may be somewhat broader or narrower than the scope of the caption, they are considered to be more-or-less coextensive with the meaning of the class. Other topics are said to be in “standing room” at the class. As topics in standing room at a class gather sufficient literary warrant, the class is subdivided to form subclasses in which former standing room topics approximate the whole of the new class. The distinction in the DDC between topics that approximate the whole of a class and topics that are in standing room in the class echoes the distinction drawn by Soergel (1985, 276) between the general references mode of a descriptor (which parallels approximating the whole) and its inclusive mode (which parallels approximating the whole plus standing room).

In the DDC, all topics in class-here notes approximate the whole of the class, while all topics in including notes are in standing room in the class. (As discussed later, some topics in multi-topic captions approximate the whole of the class, while other topics in such captions are in standing room.) For example, figure 1 gives the caption, including note, and class-here note for 796.86 Fencing. Sword fighting, given in a class-here note, is considered to approximate the whole of the class, while bojutsu and kendo (Japanese

1. DDC, Dewey, Dewey Decimal Classification, and WebDewey are registered trademarks of OCLC Online Computer Library Center, Inc.
2. Interdisciplinary works cover a subject from the perspective of multiple disciplines, including the discipline where the interdisciplinary number is located. Comprehensive works cover all the components of a subject within a single discipline.
martial arts using sticks or bamboo swords), given in an including note, are specific forms of fencing (in its broadest sense) and are considered to be in standing room in this class; either of these subjects may be given its own subordinate class at a later date.

The prior statement that the Relative Index does not indicate the relationship between the topic named by an index term and the DDC class that it indexes can now be understood to mean that no record has been kept of which Relative Index terms correspond to topics that approximate the whole of the class and which correspond to topics that are only in standing room. The study reported on here represents an attempt to make such designations automatically.

**Motivation**

Why would it be useful to ascertain the nature of the relationship between an index term and a class that it indexes?

- **Standard subdivisions** (commonly recurring forms and topic aspects—for example, serials, geographic treatment, management—that apply broadly across the scheme) can be added for topics that approximate the whole, but not for those in standing room. This restriction exists to avoid complications that arise when a standing room topic is subsequently provided with its own number: Works classed while the topic is in standing room at a number should be collocated under that one number and not scattered under its various subdivisions. At the same time, not adding standard subdivisions for standing room topics maintains the integrity of the class number plus standard subdivision notation for more general topics. Explicit indication whether Relative Index terms correspond to topics that approximate the whole or that are in standing room would help human classifiers use the scheme more faithfully and consistently. As the scheme is increasingly used in automated contexts, such a characterization could also be used in verifying the correctness of built numbers (numbers constructed by appending notation—for example, standard subdivisions—to a base number) or in computer-aided generation of built numbers.

- The process of determining whether or not an index term approximates the whole of the class it indexes begins with determining which part of a schedule or table entry the index term corresponds to. A side benefit of this determination would be to help coordinate the maintenance of the class / index term relationship: When a concept is moved from one class to another, the corresponding index terms should move with it.

- At the same time, it is known that the relationship between some Relative Index terms and their corresponding classes cannot be determined without consulting outside sources. For example, the Relative Index term “Pipil language” indexes notation 97452 Aztecan languages in Table 6, which includes only a class-here note for Nahuatl (Aztec). According to *Ethnologue* (2005), Pipil is a nearly extinct Aztecan language and can be understood to be in standing room in this class. A recommendation has been made to record the status of all such Relative Index terms (Mitchell, 2006). But such terms can only be identified by first determining for which RI terms the correspondence between term and schedule can be handled automatically.

**Figure 1. DDC class example**

796.86 Fencing
Including bojutsu, kendo
Class here sword fighting
• Addition of DDC numbers to subject authority (e.g., Library of Congress subject heading [LCSH]) records depends on the conceptual relationship between the subject heading and the DDC class(es) with which it commonly co-occurs. When Library of Congress subject headings are mapped to DDC classes, corresponding Relative Index terms are also added. The status of the Relative Index term migrates to the LCSH it corresponds to.

• We intend that rigorous analysis of the classification will eventually lead to supporting automated reasoning within the scheme, for example, determining the degree to which two developments within the scheme parallel one another. It will at times be crucial in such reasoning to know the scope of topics associated with the class relative to the class itself; do they correspond to or approximate the whole of the class or are they more specific concepts?

Methodology
Why is the determination whether Relative Index terms correspond to topics that approximate the whole of the class (for example, class-here note concepts), or to standing room topics (for example, including note concepts) anything other than trivial? For example, once the possible difference in case is accounted for, where is the challenge in determining that the index terms “Bojutsu” and “Kendo” correspond to the including note concepts in 796.86? One difficulty arises from the fact that some concepts in captions approximate the whole of the class, while other concepts in captions are in standing room in the class. Another difficulty lies in the fact that individual caption, class-here, and including concepts are often not clearly set off. For instance, the including note for 133.3 Divinatory arts reads “Including fortune-telling by bones, dice, pendulum.” Here we find three concepts—fortune-telling by bones, fortune-telling by dice, and fortune-telling by pendulum—only the first of which occurs in full in the schedule. A third difficulty lies in the attempt to adopt a form for index terms that is consistent across the scheme, and if possible, also consistent with the form of entry for Library of Congress subject headings. The Relative Index terms that correspond to the fortune-telling concepts at 133.3 include Bones—divination, Dice—divination, and Pendulums—divination. Here the difficulty of detecting the appropriate relationships resides in a difference in number between “pendulum” and “pendulums,” the substitution of a synonym, divination (a morphological variant of a word in the caption for 133.3), for fortune-telling; and a reordering of the core elements of process and instrument. A fourth type of difficulty goes beyond the typical morphological, syntactic, and lexico-semantic bases of paraphrase and extends into less-regular and more pragmatically oriented relationships. Consider these other Relative Index terms for 133.3, all of which correspond to the phrase “occult methods of foretelling the future” in the class-here note: Forecasting—occultism, Predictions—occultism, and Prophecies—occultism. The relationship of forecasting, predicting, and prophesying to foretelling the future is clear to any human who understands the meaning of the terms, but they might not be listed in lexical resources as synonyms. These phrases are provided as index terms precisely because a searcher may have any of these words in mind when trying to search for the “foretelling the future” concept.

3. LCSH/DDC mappings, which are included in WebDewey, are available on the Dewey web site at http://www.oclc.org/dewey/updates/numbers/default.htm.
Two phases are undertaken in addressing the difficulties noted above. The first phase aims to isolate concepts in captions, class-here notes, and including notes and to determine which of these concepts approximate the whole of the class and which are in standing room. (Prior to this phase, captions and notes undergo minimal preprocessing: All text is converted to lower case, while formulaic phraseology—“Class here,” “Including,” “comprehensive works on,” “interdisciplinary works on”—is deleted from class-here and including notes.) The second phase works to match Relative Index terms to the concepts isolated and categorized in the first phase. All of the work reported here has been limited to terms that index classes in the schedules, but few if any differences would be expected in processing index terms for classes in the tables or for built numbers.

Distinguishing between Approximating-the-Whole and Standing-Room Topics
While captions and notes may include only a single topic, many of them incorporate multiple topics. Where possible, we wish to isolate these topics, since Relative Index terms correspond to single topics. Topic isolation makes the matching process between Relative Index terms and schedule entries cleaner, thus increasing our level of confidence in the results.

All captions and notes that incorporate multiple topics include specific marks of punctuation or the conjunction and. We use this knowledge to identify captions and notes that are known to contain only a single topic: they lack the aforementioned markers and thus require no further topic individuation processing.

The remaining captions are scanned for a set of four spaces, which separate the parts of dual headings (e.g., Systems programs Operating systems). Both captions and notes are similarly scanned for semicolons, which always act as topic separators; thus the caption “Satellites and rings; eclipses, transits, occultations” is broken down into two parts: “satellites and rings” and “eclipses, transits, occultations.” These two parts illustrate two other rules for topic isolation. One rule says that if a part includes three words, including the word and, but no commas, and acts as a topic separator; thus, “satellites and rings” is broken down (terminally) into the topics “satellites” and “rings.” The other rule says that if a part includes one or more commas, does not include the word and, and the number of words in the part is exactly one more than the number of commas, the commas act as topic separators; thus, “eclipses, transits, occultations” is broken down (terminally) into the topics “eclipses,” “transits,” and “occultations.” These rules succeed in isolating topics in many cases, but not all. For example, the rules are unable to determine if the individual topics within the caption “Seasons and zones of latitude” are “seasons” and “zones of latitude” or “seasons of latitude” and “zones of latitude.” In such a case, the multi-topic “seasons and zones of latitude” is retained as a unit.

Once the individual topics in captions and notes have been isolated to the degree possible, another set of rules is used to identify which of the topics approximate the whole of the corresponding class and which topics are in standing room in the class. Topics in notes are easily processed: all topics in class-here notes approximate the whole, while all topics in including notes are in standing room. Topics in captions, however, require further investigation.

If a caption consists of a single topic, the topic approximates the whole of the class. Generally, if a caption begins with such words as “Miscellaneous,” “Other,” or “Specific,” the corresponding number has no meaning of its own, but exists only to introduce classes for specific examples of the topic. All such caption topics are in standing room. If
a caption consists of multiple topics, either the topics name the subdivisions of the class, in which case they are in standing room at the superordinate class level, or a note should appear indicating which of the topics or topic combinations in the heading standard or other subdivisions can be added for; such topics or topic combinations approximate the whole of the class. Caption topics that cannot reliably be categorized as approximating the whole of the class or being in standing room there are maintained separately and labeled as being of unknown status.

Matching Relative Index Terms to Schedule Entries
Once topics in captions, class-here, and including notes have been isolated and categorized, the stage is almost set for the comparison of Relative Index terms to those topics. One further intervening step is needed because the form of Relative Index entries that include subheadings is unlike the form of captions and notes, where topics are expressed in a less structured, more natural-language-like, fashion. Relative Index subheadings are used to “relate” index terms to the disciplines in which they are treated (note that the disciplinary context of captions and notes is inherent in their hierarchical placement). It should also be noted that while both captions and Relative Index terms may include parenthetical terms (i.e., qualifiers), different rules govern their use, so they are unlikely to match. Well over one-half of all Relative Index terms include subheadings or parenthetical terms, so it is important to handle these in an intelligent manner.

In the case of captions, parenthetical terms and the main heading of the caption are typically related as synonyms or quasi-synonyms. In the case of Relative Index terms, however, parenthetical qualifiers disambiguate homonyms or other ambiguous terms. It is common, then, for subheadings and parenthetical qualifiers to name concepts outside the scope of the local class, while at the same time within the scope of the hierarchy above the local class; the subheadings and qualifiers give what Miksa (2006) refers to as “conceptual context.” More particularly, many subheadings name the discipline that the class is part of. If the words in the subheading are found in the captions of higher level classes, the subheading can be safely ignored for the purposes of matching the Relative Index term to topics within the class. Otherwise, subheading words are retained as part of the Relative Index term.

In comparing Relative Index terms with topics that approximate the whole of the class, topics in standing room in the class, and topics of unknown status, four types of matches have been generated. In an EXACT match, the index entry matches a caption or note topic exactly and fully, with the possible exception of mismatch on number. (Between a first and second pass of EXACT match comparison, words that end in certain patterns typical of plurals are converted to the corresponding singular forms.) In an IN match, the entire index entry occurs as a phrase within a caption or note topic. In a KEYWORD match, all of the words of the index entry occur within the caption or note topic, but order is not necessarily maintained. In a PARTIAL match, at least half of the words in the Relative Index term occur within the caption or note topic. Punctuation is ignored throughout. For any given Relative Index term, matching proceeds through a series of steps, commencing with the most strict matching. If a match is found in any given step, further comparison operations are bypassed; otherwise, comparison criteria are gradually relaxed. For example,  

---

4. Qualifiers also occur relatively frequently with geographic names, which are, for the most part, outside the scope of the present study; only classes in the regular schedules, not those in tables, have been included.
Table 1. Matching results

<table>
<thead>
<tr>
<th></th>
<th>EXACT</th>
<th>IN</th>
<th>KEYSOWD</th>
<th>PARTIAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approximates the whole</td>
<td>11335 (24.4%)</td>
<td>4209 (9.0%)</td>
<td>1241 (2.7%)</td>
<td>5189 (11.2%)</td>
</tr>
<tr>
<td>Standing room</td>
<td>3116 (6.7%)</td>
<td>4064 (8.7%)</td>
<td>952 (2.0%)</td>
<td>1513 (3.3%)</td>
</tr>
<tr>
<td>Unknown status</td>
<td>1042 (2.2%)</td>
<td>1479 (3.2%)</td>
<td>584 (1.3%)</td>
<td>1091 (2.3%)</td>
</tr>
</tbody>
</table>

common prepositions and articles are ignored for IN and KEYWORD matches, and words are represented by their stems. Stemming used the Porter (1980) stemmer and was amplified by morphological derivation relationships from WordNet (Fellbaum, 1998). After the single PARTIAL match step, a last IN and KEYWORD match step was performed in which captions were amplified by definition, scope, and variant name notes and in which index terms, captions, and notes were amplified by synonyms from WordNet.

Table 1 gives the results of the current state of processing. Approximately 74% of the index entries find a match among the topics found in the caption, including note, or class-here note; 68% of the index entries are categorized as corresponding to topics that approximate the whole of the class or are in standing room there, of which almost four of every five are based on matches of all words (excepting stopwords) in the index term. No match was found for 11146 (24.0%) of the 46509 Relative Index terms being processed. (The sum of the counts is slightly higher than the total (by .3%) because some Relative Index terms match multiple topics within the same match step.)

Future Work

Several improvements to the matching process have not yet been implemented, for example matching on terms through additional semantic relationships (e.g., antonymy, hyperonymy) and expanding the sense of hierarchy to include structural hierarchy where it does not match notational hierarchy for the purposes of processing Relative Index subheadings.

An analysis will be undertaken of topics of unknown status, to identify, if possible, other heuristics for deciding whether topics approximate the whole or are in standing room. An analysis will also be undertaken of Relative Index terms that passed through the comparison process without finding any matches, to identify, if possible, other match heuristics. Finally, an evaluation of matches will be undertaken, to ascertain the accuracy of the categorization.

Conclusion

Being able to identify correspondences between Relative Index terms and topics in the caption and including and class-here notes of a DDC class, and to designate whether they approximate the whole of the class they index or are in standing room there, has a variety of potential benefits. Coupling knowledge of the conventions of the DDC with shallow natural language processing appears to be sufficient for the task, but this conclusion must await confirmation through more rigorous evaluation.

Acknowledgments

I gratefully acknowledge the insights and careful review of members of the Dewey editorial team.
References


John DiMarco — St. John’s University, Queens, NY

Examining Bloom’s Taxonomy and Peschl’s Modes of Knowing for Classification of Learning Objects on the PBS.org/teachersource Website

Abstract
The delivery of learning objects, small chunks of stand alone core components of reusable instruction (Nugent 2005) are used extensively in k-12 environments. Learning objects are videos and animated clips which are deployed in classrooms through public television companion websites. One such site is PBS.org/teachersource. Scholars and institutional reports (Schaffer & Douglas 2004, Nugent 2005, and PBS 2003) have identified challenges with learning objects lacking reliable pedagogical and curricular context due to the need for interpretation of raw materials by teachers and students (Nugent 2005). Technical challenges in learning objects include pedagogy and context relationships, issues with metadata in digital portals, and digital rights to content. This is a study of metadata representations of learning objects. The goal of this study was to propose and apply a comparative taxonomy to classify learning objects based on Bloom’s Taxonomy in 2006 and Peschl’s modes of knowing in 2007, thus adding to the discourse on taxonomies present in learning object repositories and knowledge-oriented educational processes existing in public broadcasting and educational new media content domains. In an effort to re-examine the data and conclusions, I matched Bloom’s model against modes of knowing keywords put forth by Peschl (2006). Matching the Peschl model against the Bloom model provided a new template to use in determining the cognitive level of learning objectives embedded in PBS TeacherSource website learning objects.

1. Introduction
Teachers of k-12 students are utilizing instructional media including television and integrated web portals within the classroom increasingly due to a growth in educational content production and an increase in computers and coaxial cable and fiber optic networks in schools. The delivery of learning objects, small chunks of stand alone core components of reusable instruction (Nugent 2005) are used extensively in k-12 environments. Learning objects are videos and animated clips which are deployed in classrooms through public television connected to a companion website. The companion website provides a rich media source on the internet that integrates interactivity and games to provide an enhanced learning experience for the viewer. Technical challenges in learning objects include pedagogy and context relationships, issues with metadata in digital portals, and digital rights to content. Each technical challenge presents opportunity for further research. Specifically, the understanding of content and context relationships by teachers during retrieval of educational materials presents an interesting problem. The goals of this study are to first examine the problem of metadata representation in classifying learning objects. The second goal is to use the data gathered to define elements empirically with a taxonomy of learning objects that classify “Interactive Online/Activities” on the PBS.org/teachersource educational website. PBSKids.org/teachersource is a connected website of PBS.org that provides online learning objects used in classrooms for pre-k through high school students. Scholars (Li, Nesbit, Pesin, Specht, Richards et al. 2003, 2006) have identified and studied learning object repositories including websites, knowledge management systems, learning management systems, and a variety of e-learning frameworks. Throughout the studied deployments, researchers have identified various systems and paradigms for mapping between learning objects between, languages, ontology’s, and work practices. One common theme for
further research suggested in these studies is to build a higher degree of communication and interaction between communities that work with learning objects. These communities exist on many levels, in many different training and learning contexts. The focus of this study is to look at the k-12 user website content to understand and organize metadata representations of learning objects associated into a useful taxonomy that can yield more effective information retrieval models across educational contexts, which can then foster increased communication and interaction for teachers. To do this, Bloom’s Taxonomy will be used as a template for collocation and categorization of learning objects at the PBS.org/teachersource website.

Bloom initiated the idea that developing a framework would be a good means for exchanging test items among faculty at various universities through creating banks of items with the same objectives. Bloom utilized a group of experienced measurement specialists from across the United States, and met about twice a year beginning in 1949 to consider progress, make revisions, and plan the next steps (1). Their final draft of the curriculum research was published in 1956 under the title, *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain* (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956). This work is referred to as the original taxonomy and was used to generate the template for this taxonomy development project. The Bloom template was adapted using Peschl’s modes of knowing with the goal of re-examining the data against a modified template. The Peschl model (2006, 63) is based on the realization that knowledge and knowledge constructions are considered to be the heart of educational and knowledge sharing processes, which connects with the Bloom model, but brings a new caveat; each level depends on each other. This assertion by Peschl was used to re-examine the data set collected in December 2006 against Peschl’s modes of knowing as it sits in the framework of Bloom’s Taxonomy.

The purpose of this research is to contribute to evolving discourse within the educational media community on classification of learning objects. Specifically, the understanding of content and context relationships by teachers during retrieval of educational materials presents an interesting problem. K-12 teachers need to access educational materials inside and outside class time. The educational materials that need classification are learning objects residing at the PBS.org/teachersource internet site.

Looking at learning objects through the lens of Benjamin Bloom’s Taxonomy of Educational Objectives for the cognitive domain might help contextualize materials better for information retrieval by teachers in classroom and curriculum development experiences. Bloom (1956) ordered his taxonomy and the whole cognitive domain so that each classification level within the cognitive domain required the skills and abilities of the items lower in the order. Bloom’s taxonomy (1956) of educational objectives is defined for this study based upon their original definitions put forth by Bloom (editor) and others (Engelhart, Furst, Hill, & Krathwohl). The brief definitions of each domain are as follows:

- **Knowledge** which includes behaviors or test situations which emphasize remembering of ideas, material, or phenomena, is the most basic level of the domain.
- **Comprehension** which includes objectives, behaviors, or responses which represent an understanding of the literal message contained in a communication, tests or measures translation, interpretation, and extrapolation of knowledge.
**Application** includes the objective that tests the person’s ability to use knowledge correctly without being given a prior solution. This goes beyond comprehension which translates and interprets knowledge, but does not apply it.

**Analysis** includes breaking down of materials into parts and detecting relationships and organization of the parts for establishing meanings or conclusions of communications.

**Synthesis** includes working with elements or parts to combine them in order to form new patterns or structures that did not exist before.

**Evaluation** includes making judgments about the value of ideas, works, solutions, methods, materials and processes. Evaluation represents the most advanced level and requires the combination of all the other five cognitive behavior levels.

Developing a clear taxonomy for K-12 learning objects that utilizes Bloom’s Taxonomy in developing a new taxonomy will have value as the growth of learning objects causes new works to emerge and existing works are instantiated and repurposed in multiple media including VHS, DVD, and web archives.

The need for understanding has yielded research on the evaluation of learning objects. Leacock (et al. 2004), Li, Nesbit, and Richards (2006) have provided data on web based learning object evaluation tools such as eLera, a network of interoperable Canadian repositories launched in 2003 and federally funded. eLera stores metadata pointers to learning objects (2006) and does not store the objects themselves, as a website such as PB-Skills.org would. eLera provides “quality ratings and comments in the form of metadata that can powerfully assist in the search and retrieval of objects” (19) and uses LORI, the Learning Object Review Instrument (Nesbit, Belfer and Leacock 2003), which allows content reviewers to enter comments and ratings on a five point scale. The main conclusion across the learning object evaluation literature is that methods for recommending objects and mapping metadata and reviews can lead to increased communication and interaction between communities that work with learning objects (Li, Nesbit, Richards 2006).

Evaluation of learning objects research has value in solving the problem of developing a taxonomy of learning objects to address the central problems of increasing communication and interaction for the benefit of user retrieval of learning objects.

Learning objects are delivered through systems that utilize data repositories. Douglas and Shaffer (2004, 15) suggest that: “[the] purpose of repositories is to support problem solvers, designers, or learners by providing a centralized location for the storage and reuse of standard artifacts and objects.” Artifacts are templates and other documentation used during the analysis and design phase of the learning object development process. Common communities and developing taxonomies structured towards those communities for specific information retrieval appears in the literature as a major concern in learning object research in repositories and classification. Research suggests that classification of learning objects is complicated. Li, Nesbit, and Richards (2006, 27) suggest that the complexities are due to an abundance of criteria that can be selected and called for employment of mapping a learning object to more than one classification scheme and suggest that rich taxonomical information indicating the local learning outcomes, local content descriptions, and pedagogical design would enrich object descriptions.

User-centered research in classification of content for image retrieval has been examined by many scholars across multiple disciplines due to the wide variety of systems and contexts that take the form of information retrieval. Marsh and White (2003) explored the relationship of images and text to understand the implications of the research for in-
taxonomy that allows user descriptions to be mapped onto interface design solutions for
image databases and image retrieval systems. Pescin, Specht, and Adam (2003, 1) dis-
cussed “how to create efficient and concise representation or indices of the digitized data
to facilitate search and retrieval” and the need for a flexible approach to managing learn-
ing object metadata in learning content management systems (LCMS), which provide a
retrieval mechanisms for learning objects.

Universal taxonomy of learning objects is problematic due to the context sensitive na-
ture of learning objects and the disparate nature of the information retrieval systems that
deliver learning objects. Clyde (2004) points out learning objects are hard for teachers to
find and utilize learning objects due to the lack of classification and taxonomy standard-
ization in learning object systems. PBS.org/teachersource uses standards based and grade
level based taxonomy. This approach is weak because it does not bring the learning object
information seeking process deep enough into instructional objective contexts. The uni-
versal concepts put forth in Bloom’s taxonomy are commonly known to college educated
teachers due to the relevance they bring to curriculum and instruction and the inclusion of
the topic in curriculum development courses taken by student teachers. Therefore, using
Bloom’s Taxonomy in the development of a new taxonomy for learning objects brings a
standardization of sorts, which directly relates to the context of the k-12 classroom teach-
er. Exploring Bloom’s Taxonomy to bring familiar context to learning object retrievals
reveals an interesting research question: What cognitive domain categories of k-12 learn-
ing objects emerge from content analysis of learning object descriptions on the PBS.org/
teachersource website using Bloom’s Taxonomy as a framework for classification? Mov-
ing Bloom’s model forward to connect it with Peschl’s modes of knowing allows another
direction for discourse, one that brings an educational model and an information studies
model together for data analysis. Which categories emerge against Peschl’s modes of
knowing as they are compared against the framework of Bloom’ Taxonomy?

2. Peschl’s modes of knowing and knowledge
Peschl (2006) describes levels of knowledge as they intersect with process/activity, and
resulting knowledge. The identification of the levels of knowledge, modes of knowing,
and the cognitive activities necessary for developing these modes is presented by Peschl
for the sake of improving learning/teaching processes according to knowledge creation
and knowledge structures. Peschl presents five different levels that are reliant on each
other and build from the behavioral level (1) to the level of reflection (5) (64). These
levels seem to logically connect to Bloom’s six levels of learning objectives in the cog-
nitive domain. The behavioral level, which is resulting knowledge, is connected to the
first level of Bloom’s taxonomy, knowledge. The second level presented by Peschl is the
level of emerging patterns and behaviors, which is characterized by explanation. This
level is connected to Bloom’s second level, comprehension, which objectifies translation.
Peschl’s third level, the level of cause and the source, breaks sequence as it relates to the
fourth level of Bloom’s taxonomy, analysis. The fourth level of Peschl’s modes of know-
ing is the level of potentiality, change, and design. Level 4 is characterized by exploring,
developing, and design, which connects it to both application (level 3 of Bloom’s taxon-
omy) and synthesis (level 5 of Bloom’s taxonomy). The final level of Peschl’s modes of
knowing, level of reflection (of the causes, source, patterns, and processes of knowledge
construction) connects to Bloom’s level six, evaluation.
### Figure 1. Levels of knowledge put forth by Peschl and adapted to Bloom’s Taxonomy*

<table>
<thead>
<tr>
<th>1.0 Behavioral level</th>
<th>2.0 Level of (Emerging) Patterns and Behaviors and Relationships</th>
<th>3.0 Level of Causes and the Source (Analysis, coded as C)</th>
<th>4.0 Level of potentiality, change, and design (Application, coded as D and Synthesis)</th>
<th>4.0 Level of potentiality, change, and design (Synthesis, coded as E)</th>
<th>5.0 Level of reflection of the causes, source, patterns, processes of knowledge construction (Evaluation, coded as F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Knowledge, coded as A)</td>
<td>(Comprehension, coded as B)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The first levels of Bloom are directly related to Peschl’s level 1 and 2. Level 3 and Level 5 of Bloom is related to Level 4 of Peschl. Level six of Bloom is related to Level 5 of Peschl.

### 3. Methodology

The taxonomy was developed in two stages. The first stage used Bloom’s Taxonomy to generate exhaustive keywords describing each of the six levels for a total of 140 outcome illustrating verb keywords. The original list from Krathwohl was analyzed against Krumme’s list of outcome verbs (2005) and converted into a keyword list. The terms were derived and a taxonomy table (Figure 1) was developed and used as a template to analyze search descriptions of 322 learning object descriptions gathered on December 5, 2006 from the PBS.org/teachersource website. In the second stage, I collated and identified concepts from cognitive based learning categories extracted from search descriptions of learning objects from a PBS.org/teachersource search. After stages one and two were complete, the research was triangulated by adding a new model of inference, Peschl’s level of knowledge and modes of knowing. The data were examined to reveal new conclusions and to answer the research question: Which categories emerge against Peschl’s modes of knowing as they are compared against the framework of Bloom?

### 4. Data Analysis

Each X column of Bloom’s Objectives were coded using sequential letters with A representing Knowledge, B representing Comprehension, C representing Application, D representing Analysis, E representing Synthesis, and F representing evaluation. Classification of nominal data was used to determine answers to the research question; what categories of k-12 learning objects emerge from content analysis of learning object descriptions on the PBS.org/teachersource website? All combinations of collocation data were measured nominally to count total occurrences of keywords and outcome illustrating verbs. Each search result content description was analyzed against the frameworks published by Krumme (2005) and Krathwohl (2002). Coding for each description required identification, template matching, and collocation of learning objectives snippets matching keywords and outcome illustrating verbs. Once matched, the snippets were placed in the appropriate cell of the Taxonomy Table and then coded for representation in one of the six levels of learning objectives in Bloom’s Taxonomy.

Peschl’s levels of knowledge were set against Bloom’s taxonomy template to see if there was a connection between the two models. Connections were made by matching words and descriptions of each level of Bloom to a Level of Peschl. By connecting Peschl’s work to Bloom’s work, similarities between the two were revealed.
5. Conclusions
Several interesting conclusions can be drawn from the data sets after using Bloom’s Taxonomy as a template for learning object classification:
1. The search for “InteractiveOnline/Activities” at PBS.org/teachersource revealed 322 records, each of which contained a text snippet describing the learning objective(s) of the online learning object. After content analysis, all occurrences of learning objectives matching Bloom’s Taxonomy totaled 616.
2. Knowledge, the lowest level (1) of educational objective of Bloom’s Taxonomy, had the most evidence of occurrence in the PBS.org/teachersource learning object population at 27.1%.
3. Level 3, Application, in single form ranked second in total occurrences with 12.8%.
4. Forty combinations of learning objects emerged, of which the highest occurrences were knowledge and application (Peschl’s Level 1.0 and 4.0).

These conclusions answer the research question: What cognitive domain categories of k-12 learning objects emerge from content analysis of learning object descriptions on the PBS.org/teachersource website using Bloom’s Taxonomy as a framework for classification? And, by showing the emerging categories, we see that the learning objects at PBS.org/teachersource are instructionally designed on achieving objectives that are lower in the cognitive domain based on the research of Bloom and his editorial group and the subsequent instantiations of the 1956 work in 2002 by Krathwol and in 2005 by Krumme. As well, Peschl’s level’s of knowledge has similarities in words and meanings that provide an additional source for interpretation of cognitive level of learning object on the PBS TeacherSource website. Currently, PBS.org uses a standards based taxonomy for search retrieval of learning objects. The problem with this classification structure is that it does not provide a contextual based taxonomy that is focused on learning objectives for retrieving learning objects. Utilizing Bloom’s familiar taxonomy and contextual lens, the information seeking activities of teachers looking for meaningful learning objects can go beyond baseline grade level headings or subject headings. Placing learning objects in a familiar objectives-based, contextual framework such as Bloom’s Taxonomy or by comparing it against another domain specific model, such as Peschl’s levels of knowledge, as demonstrated in this study, might help instructional designers, and teachers accelerate success in transferring learning objectives using learning objects. Further studies in learning object taxonomies are suggested to build scientific literature to help foster increased research into the highly varied communities that rely on learning objects for teaching and learning. These communities include pre-k through 12 schools, higher education, training, and distance learning. Exploration of different taxonomies for varied communities is also suggested as future research to increase the discourse on taxonomies of learning objects in web based applications. Additionally, research is needed to determine why learning objects may sit lower on the cognitive level, perhaps due to the need for usability or lack of instructional objectives during the design process.

Acknowledgement
I would like to sincerely thank Dr. Richard Smiraglia for his mentoring and expertise.
References
Melanie Feinberg — University of Washington

Classificationist as Author
The Case of the Prelinger Library

Abstract
Within information science, neutrality and objectivity have been standard design goals for knowledge organization schemes; designers have seen themselves as compilers, rather than as authors or creators. The organization of resources in the Prelinger Library in San Francisco, however, shows a distinct authorial voice, or unique sense of expression and vision. This voice, in turn, works as a persuasive mechanism, facilitating a rhetorical purpose for the collection.

Introduction
According to standard methods, the task of designing an organizational system, be it a classification, controlled vocabulary, or metadata schema, should be approached with detachment and objectivity. Even as classification researchers acknowledge, as does Clare Beghtol (2001), that “every classification is a theoretical construct imposed on ‘reality,’” the classificationist is still seen as someone who compiles, or documents, the perspective of a defined group or groups (perhaps that of a particular discourse community, organization, or other set of users). This standpoint is sensible when considering professionally developed classifications created to facilitate the retrieval of documents for some defined public. It would not be useful, in such a scenario, for a classification to exhibit the original perspective of its creator.

The Prelinger Library in San Francisco provides a counter-example to the idea of classification as documentation. While the Prelinger Library provides public access to its materials, it does not operate under a standard retrieval orientation. (Megan Shaw Prelinger describes the library as being “browsing-based” rather than “query-based,” to “[open] wide the possibility of discovery.”) In this paper, I explore how, in violating standard classification design goals of neutrality and predictability, the Prelinger Library’s classification system shows an authorial voice, or vision. This voice, in turn, facilitates a rhetorical purpose for the classification.

For the Prelinger Library, authorial voice as represented in the classification system is tightly integrated with the content of the collection. This again goes against professional practice, as selection of resources is often regarded as a separate task from their representation. While the Prelinger Library comprises a discrete set of physical resources housed in a particular location, this design strategy is also apparent in a common feature of social classification systems: the sharing of personal resource collections. The paper concludes by suggesting that authorial voice, as expressed through the combination of selection, description, and arrangement, might be a useful construct in both understanding and designing such shared collections.

Classificationist as careful compiler: the traditional view
Neutrality has been a persistent goal in classification design. The necessity of neutrality in nomenclature is Sayers’s tenth “canon,” or postulate, for classification design (Sayers, 1915). “The introduction of any name which exhibits a critical view of the subject it connotes is a violation of one of the first principles of classification,” Sayers admonishes (Sayers, 1915, 32). While scholars have debated at length the scope of what might
be accurately documented (the whole of knowledge, a single subject field, the multiple perspectives that constitute a particular discourse community), the sense of the classificationist as someone who uncovers an existing order, as opposed to creating a new order, seems fairly consistent. Ranganathan’s canon of helpful sequence, for example, seems to function as a logical ideal, a form of external reality that the classificationist should attempt to isolate, and that the classification should faithfully reflect; there is a most helpful sequence, and Ranganathan’s set of canons, postulates, and so on, enables its attainment (Ranganathan, 1959). Similarly, in contending that classifications should “adapt […] to the existing structure of thought,” Shera makes the identification and representation of this structure the classificationist’s goal (Shera, 1966, 84). Shera’s “existing structure of thought” seems similar to Beghtol’s description of “cultural warrant,” a term that she uses to encapsulate the changing meanings of literary, scientific/philosophical, and educational warrant over time (Beghtol, 1986). If neutrality is attained and the chosen scope well documented, then the classification should be predictable for the selected user group, and thus useful in a retrieval context. It is precisely the failure of classifications such as the Dewey Decimal Classification and Soviet library classification to achieve these goals that motivates Clay Shirky’s attack on all forms of “ontology” as “overrated” (Shirky, 2005).

While Andersen (2000) suggests that an indexer should be considered an author, and that an indexer, in determining the descriptors to assign to a document representation, should consider the four contexts of writing described in Bazerman’s (1988) discussion of scholarly articles (the object of study, the literature of the field, the audience, and the author’s own persona), it is not clear from Andersen’s analysis how document representation might change with such an approach (if, for example, neutrality would no longer be a goal, and if so, how the absence of neutrality would then affect the experience of using an index).

**Authorial voice**

Although intuitively understood to mean a sense of the author’s presence and vision, voice is a tricky, even controversial, concept. Peter Elbow (2007) describes how concepts of voice have changed within the field of composition studies, or the teaching of writing: in the 1960s and 1970s, the idea of voice was associated with a sense of agency, authenticity, and “rhetorical power,” in keeping with a pedagogical focus on individual expression. However, in the 1980s and 1990s, voice, and “individualist” pedagogy in general, was subjected to postmodern critique. Scholars and writing teachers focused more on social construction of the text, in particular on the constraints of dominant discourse structures, and became cynical about authorial intention, which seemed inextricably connected with voice. The idea of the writer as a coherent self or any sort of autonomous agent was repudiated as an artifact of the hegemonic discourses of capitalism and patriarchy (Faigley, 1992).

However, one need not equate the expressive qualities of a text with a particular author’s “true” identity or intentions. Wayne Booth (1983) asserts that a sense of authenticity in writing results from the textual construction of an “implied author,” not from the actual self of the writer. In addition, even within the coordinated structures of a particular discourse community, rhetorical situation, and set of genre conventions, writers are able to make choices that contribute to a sense of authorial presence (Clark and Ivanic, 1997; Johnstone, 1996). Johnstone, a sociolinguist, comments that “self-expression plays a crucial role in […] mediating between options and outcome” and notes as well that “even the
most formulaic genres,” such as thank-you notes, can be “self-expressive in the hands of good writers” (Johnstone, 1996, 90 and 179). In this vein, Matsuda defines voice as an “amalgamative effect of the use of discursive and non-discursive features that language users choose, deliberately or otherwise, from socially available yet ever-changing repertoires” (Matsuda, 2001, 40). Voice, in this definition, represents the overarching quality of difference that distinguishes one writer’s response from another’s, despite the inevitability of various social constraints. In fact, one could say that such constraints contribute to the experience of voice by marking deviations. A thank-you note that manages to express a singular voice will indeed stand out.

The textual features that combine to convey voice are often described as stylistic, referring to form but not content (as in Elbow, 2007). Clark and Ivanic (1997) comment, however, that what is being said may reveal authorial presence more strongly than the manner in which content is presented. An author who claims ownership over unique ideas (as opposed to, for example, primarily citing the ideas of others) has more presence in a text, even if the style is unobtrusive. This detail is particularly important for the investigation of voice in classificatory texts, which lack the range of stylistic resources possible in more conventional forms of writing. We may find voice not merely in the nomenclature used for categories or other primarily expressive techniques, but also in the constitution of classes and their relationships, and in the assignment of categories to selected resources. The way that classes are defined and used, in addition to the way that they are named, may show the confluence of imagination and vision that the concept of voice represents.

**Voice in the Prelinger Library**

The Prelinger Library is a non-circulating private institution with a collection of 50,000 items. The items are not catalogued, but they are arranged in a progressive order from one end of the library to the other, and different sections of the shelves are physically marked with subject headings written on masking tape (for example, a series of headings on shelf 5 runs from U.S. Internal Dissent to Nuclear Threat, then to War, Conflict, and on to Peace, followed by Radical Studies and then Utopia). In contrast to the standard design goals of neutrality and predictability, the Prelinger Library’s classification shows personality and surprise. These characteristics combine to endow the Prelinger collection with an authorial voice.

In an online essay, Megan Shaw Prelinger describes the library’s organization as a conscious attempt to “represent the realms of thought that bounce around the insides of both our [Shaw Prelinger and her husband, Rick Prelinger] minds” in a coherent linear flow across the library’s six shelves. The library’s primary organizing principle, location, provides one example of how this personality is expressed through the collection and its organization. When location is relevant, resources are classified according to location over subject in most cases. Moreover, classes based on location (primarily an idiosyncratic selection of U.S. states, regions, and cities; examples include Pennsylvania, Tennessee Valley Authority, California, and San Francisco) are placed first in the sequence, suggesting to the browser that, although the Prelinger Library may physically resemble typical libraries with standard classification schemes in some ways, this library is also significantly different from one’s local public branch. In other words, although the Prelinger Library claims membership in the familiar genre category of the library, it also adapts certain genre conventions innovatively.
In a magazine profile of the library, Lewis-Kraus ties the location principle to the Prelingers’ personal outlook, claiming that “landscape anchors not only the library but the Prelingers’ own approach to most intellectual questions” (Lewis-Kraus, 2007, 50). This quotation is testament to the library’s success in conveying an authentic sense of voice. Note that it doesn’t matter if the library’s location principle really does represent the Prelingers’ thought processes accurately; the point is that an outside interpreter believes that the library itself is expressive of a particular worldview and personality. The location principle presents a convincing vision that obtains through multiple texts: Prelinger’s online essay, Lewis-Kraus’s commentary, the classified collection itself.

As another example, the Prelinger Library’s classification has a singular emphasis on the concept of infrastructure, which runs throughout the succession of classes. Categories related to production and resulting products are followed by categories related to services on which those production processes rely. Categories for manufacturing and industry are followed by categories for transportation, power systems, and urban infrastructure such as sewers. Media products and production (television, film, video, radio) are followed by broadcasting and communications infrastructure, including telephone, telegraph, and computer networks. The repetition of this ordering and the detail with which infrastructure technologies (including social technologies, such as urban planning) are enumerated suggests a political concern, an exhortation not to forget the complex variety of systems upon which our production processes depend. Here, the sensation of authorial voice comes from the unfamiliar juxtaposition of these blunt political ideas as embodied within a seemingly conservative and standard set of genre conventions (the arrangement of books in a library).

In addition, the selection and distribution of resources help to shape both the organization of the library and the user’s experience, and particularly facilitate a sense of surprise. The collection comprises many forgotten publications, substantially from the first half of the twentieth century, most of which would initially seem to have a short “shelf life” (examples: Practice and Science of Standard Barbering, from 1951; A Study of Cider Making in France, Germany, and England, with Comments and Comparisons on American Work, from 1903 (a government-sponsored work); Report on a City Plan for the Municipalities of Oakland and Berkeley, from 1915; Big Dam Foolishness, The Problem of Modern Flood Control and Water Storage, from 1954). There is little contemporary or popular material. Runs of old serials, primarily on industrially focused topics, have been incorporated throughout; most of these are castoffs pruned from other libraries (examples: Bus Transportation, Candy Manufacturing, Modern Plastics, Retail Lumberman, Texas Police Journal). Surprise is also achieved by interleaving ephemera (often in separate boxes) within the book shelves. The transportation section, for example, includes, in addition to books about rail travel, a shelf of nineteenth and early twentieth century local train schedules from various parts of the United States. The section on domestic environments includes an extensive number of advertisements for household appliances from about 1900 to 1960 (example: “How Ironrite Freed Me from My Hardest Home Drudgery: Hand Ironing!”).

**Through voice, a rhetorical purpose**

Through the construct of authorial voice, the interwoven processes of selection, description, and arrangement provide evidence of a rhetorical motive in the Prelinger Library. The basic rhetorical process described by the literary critic Kenneth Burke via a meta-
phor of courtship appears to be at work (Burke, 1969). According to Burke’s courtship model, an author (or rhetor) first entices the reader (or audience) by emphasizing the essential differences between rhetor and audience (heightening the “mystery”) and then, as the audience’s attention is engaged, by showing how the audience and rhetor, despite their divisions, also share deep similarities (such as working for the same goal or other characteristics), resulting in identification between the rhetor and audience. As Burke says, “In mystery there must be strangeness; but the estranged must also be thought of as in some way capable of communion. There is mystery in an animal’s eyes when a man feels that he and the animal understand each other in some inexpressible fashion” (Burke, 1969, 115).

Identification, for Burke, represents the primary mechanism of rhetoric, the means by which the rhetor ensures the cooperation of the audience, as well as the goal of rhetoric. When identification has been achieved, the audience feels as if it is collaborating in the opinion voiced by the rhetor, that audience and rhetor are united in the same purpose; they are, in a sense, consubstantial. Burke elaborates that:

A is not identical with his colleague, B. But insofar as their interests are joined, A is identified with B. Or he may identify himself with B even when their interests are not joined, if he assumes they are, or is persuaded to believe so… In being identified with B, A is “substantially one” with a person other than himself. (Burke, 1969, 20–21)

Burke additionally associates identification with imagination: in referring to the work of Hazlitt, Burke claims that imagination creates possibilities that, as led through identification, an audience may desire to enact or avoid. When Burke contends that “the poetic house is built of identifications,” he is noting how a skilled writer can evoke a cluster of associations with a single well-chosen image (Burke, 1969, 85). Depending on how it is invoked, the poetic image of a house can evoke identifications with childhood, security, prosperity, and so forth, often all at once. The image thus works on a variety of levels to bring the author and audience closer together.

In the case of the Prelinger Library, “mystery” is evoked on one level by the initial strangeness at seeing thousands of overtly mundane publications within the seemingly familiar structure of library shelves, then compounded through the unusual organization of these seemingly oddly chosen resources. The initial presentation of these “useless” items puzzles the user and sets up the sense of division. The authorial voice, though, as manifested in the selection, description, and arrangement, suggests that, on the contrary, these apparently worthless items deserve preservation and care. These items have not merely been warehoused, as they might be in a used bookstore or library fire sale; they have been consciously gathered and painstakingly arranged for a very particular experience of access. This sense of care provides the pivot point for an identification to emerge. The visitor to the Prelinger, even if not charmed by old train schedules and the like, identifies with the affection and effort lavished by the Prelingers on their collection. Together, the collection and its classification suggest that all information, however negligible as it may initially seem, deserves preservation, and the visitor is persuaded to give the library’s contents more serious attention. This identification is imaginatively reinforced by the infrastructure emphasis in the library’s collection: the idea that the products upon which we rely depend themselves on often-ignored infrastructure. Information requires infrastructure for both preservation and access; the library provides this infrastructure.
The new bibliography: communicative classification

While it may be said that a classified collection, the entity that most users actually experience, is the primary carrier of meaning, as opposed to the classification itself, most professional classifications, unlike the Prelinger Library’s, are not designed with a specific collection in mind. However, the activities of selection, description, and arrangement are tightly coupled in the shared document collections enabled by various social classification systems (such as del.icio.us, LibraryThing, and Flickr). A tag, for example, does not exist in del.icio.us without being attached to a document.

Hendry and Carlyle (2006) claim that Internet-based shared collections can be seen as a new form of bibliography and suggest that bibliography might provide a conceptual base for such systems. However, bibliographic handbooks provide few details on the selection activity and how it might intersect with arrangement (as, for example, Robinson, 1979, who defers the selection task to “experts”). The only truly acceptable principle for selection often seems to be completeness. For example, the historians of bibliography Besterman and Balsamo both admire the sixteenth-century bibliographer Conrad Gesner, who attempts to compile and describe all printed works, and disapprove of sixteenth-century Catholic bibliographers who created selective works based on church doctrine. Balsamo and Besterman imply that any selection principle other than comprehensiveness is ethically irresponsible (Besterman, 1936; Balsamo, 1983). Balsamo, for example, describes Gesner’s approach of including all existing works and providing commentary on them as “without precedent in fullness and accuracy,” and notes how similarities between Gesner’s concerns and those of modern cataloging and bibliography “confirm the universality of the methodological solution happily achieved by Gesner” (Balsamo, 1983, 34 and 41) On the other hand, the bibliography of the Jesuit author Possevino, who is set up by Balsamo as the “anti-Gesner,” is a “proscriptive bibliographic canon which would serve as a tool for imposing ideologically correct works” and “a total cultural program, one without alternative […] issuing from a single dogmatic mold, with no provision for individual choice.” Possevino’s “rejection of Gesner’s classification scheme” in favor of a theologically based system of organization and selection is described as “the affirmation of a totalitarian vision which denied the autonomy of human knowledge” (Balsamo, 1983, 47).

While Bates (1976) does grant the inevitable selectivity of bibliography, she focuses on acknowledging selection principles at work, and not on studying them to see what they contribute to a bibliography’s interpretation of the subject, as manifested in the Prelinger Library through the construct of voice. The example of the Prelinger Library shows that the integration of selection, description, and arrangement may be a key element in formulating authorial voice, and, further, that this voice can partially constitute an experience of access that is different from retrieval-focused systems of organization. While it may be, as Hendry and Carlyle (2006) suggest, that bibliography provides a convenient frame to examine organizational schemes in which selection, description, and arrangement intertwine, such as social classification systems and the Prelinger Library, it seems like bibliography itself requires some expansion to account for phenomena such as authorial voice. Given the ever-increasing popularity of social classification and other forms of expression that center on the citation and organization of resources (such as iTunes playlists and even Facebook friends), renewed scholarly attention to systematic bibliography, particularly in its potentiality for creative expression, certainly seems worthwhile.
References
Hermeneutic Approaches in Knowledge Organization
An Analysis of Their Possible Value

Abstract
This paper considers how hermeneutics and other related theories may bring new insights into KO. They provide a most realistic representation of the complexity of knowledge and meaning according to which new forms of KOSs could be designed. Computational and conceptual aspects of these issues are discussed taking into account a number of case studies.

Introduction
The development of knowledge organization (KO) has strongly been influenced by logical positivism. However, an increasing attention to other theoretical approaches, and in particular to hermeneutics, can be found in many studies in KO and related fields (Chalmers, 2004; Cole & Avison, 2007; Fonseca & Martin, 2005; Hansson 2004; Hjørland, 2007). With its basic idea of the ‘situatedness’ of understanding, contemporary hermeneutics has the potential to become a theoretical background on the basis of which new forms of knowledge organization systems (KOSs) are conceived. Wittgenstein’s late philosophy, which shares some similarities with parts of hermeneutics, is also considered for such a purpose. The possibility of applying these approaches to KO issues is then investigated, both in their computational and conceptual aspects, by analyzing some case studies.

1. Towards hermeneutic approaches in KO: computational applications
1.1 The heuristic approach for the development of IS methodology
Information systems (IS) using thesauri, classifications and other KOSs adopt from their beginnings a heuristic model of reasoning deeply influenced by logical positivism. This includes, for example, attempts to codify human reality into formal definitions. Accordingly, representations of phenomena are fixed and based on logics and exclude representations of individuals’ context and situation (Chalmers, 2004).

The concept of heuristics refers to a reasoning that advances learning in basic principles to progress in knowledge through successive stages, as to enable the accumulation of successes and learning from errors. The reasoning counts on fundamental concepts that enable coding the well-known. This codification provides the possibility of extending, strengthening, or revealing it, thus creating the foundations of new knowledge in relation to the already known.

The development of IS with a heuristic model follow two knowledge representation methodologies: the Procedural form and the Declarative form. In the former, knowledge is integrated into the computer program. The knowledge representation is independent from the computational process: the computer program is seen as a collection of functions or instructions given to the computer. In the latter, a computer program is as a set of interacting individual units or objects, where any object can manage its own state and operations. Both methodologies, although implying different implementation systems, have three logical steps in common:
1. Conceptual scheme construction: representation of the system domain in terms of identifiable objects or elements with their properties and attributes.

2. Enunciation of general rules of behavior and listing of the relationships between the objects or elements of the conceptual scheme applicable to the situations upon which the system sets out to act.

3. Rule application in order to get to conclusions and modifications induced by successes and errors.

These logical steps presuppose the concepts of essence, success and truth, as well as an ideal construction of the reality upon which the IS operates (Bosch, 2006).

The heuristic approach has enabled very important advances in information systems. However, at present heuristic models are finding it very difficult to maintain their consistency and stability, since the resources to which heuristic analysis is applied include non-textual materials, multilingual environments, and important cultural differences in user profiles. In addition, the subject structures of KOSs call for multidisciplinary knowledge organization.

1.2 Towards hermeneutic approaches: spoken prose without knowing it

The hermeneutic approach is mainly rooted in Gadamer’s philosophy (1976), who insists on the ‘situatedness’ and historicity of human understanding and its priority as an interpretative and dialogic activity. Meaning is inter-subjectively created and acquired through interpretation experiences. Yet, if our understanding always depends on a given fore-structure, acts of interpretation can themselves contribute to changing this structure, in a process described as the hermeneutic circle (Chalmers, 2004).

The information in IS development with a hermeneutic approach is not merely data and facts, but also representations built through interpretation and dialogue supplying meaning and context to actions (Winograd & Flores, 1986). Like heuristic models, they use conceptual schemes. But these are not necessarily hierarchical and can instead be cross-sectional or integrated, featuring simultaneously different structures.

With the advent of the Web 2.0, various types of circulation of meaning, supported by partial, individual, and sectorial interpretations, have arisen. For example, the user is no longer one that only uses: he now becomes a protagonist. As a negative consequence, this new situation has brought about an explosion of new forms of knowledge, with overlaps in a jungle of labels. At the same time, it has also given room to innovation by and for users. This innovation comes not from the experts, but from people with investigatory interest, who use individual and partial forms of reasoning (Von Hippel, 2007).

Like Moliere’s Bourgeois Gentilhomme, the application experiences of the hermeneutic approach in IS surpass the explicitly enunciated theory. It is therefore necessary to identify the major characteristics of the hermeneutic and the heuristic approach in IS and then to investigate how they possibly apply to concrete experiences.

1.3 Heuristics and hermeneutics features of ISs using KOSs

To the main features of the heuristic approach belong: hierarchical organization; separation of developers and users; closed systems and being strongly result-oriented.

The hermeneutic approach, instead, includes primarily: pragmatic forms of knowledge classification; contextual aspects of application domain; innovative experiences of software engineering; open systems with strong user participation; and tendency to collaborative construction of knowledge.
1.3.1 A case study based on the analysis of knowledge generators

An analysis of heuristic and hermeneutic features in computer applications of KOSs has been made on selected samples from the European area, especially Italy. The knowledge generators have been chosen as dimension of analysis. We have considered the use of IT in experiences of collaborative construction of knowledge. Following the stages of the hermeneutic circle (fig. 1), we have tried to explicate our pre-understandings and established a dialog with colleagues, leading to reflection and reconstruction. The result is a scheme suggesting an approach according to types of organizations and technologies, their participants, ways of interacting and surroundings.

![Image of the hermeneutic circle as a practical research framework](image)

**Table 1. Comparative scheme with possible approach**

<table>
<thead>
<tr>
<th>Organizational modality</th>
<th>Participants</th>
<th>Interaction</th>
<th>Surroundings</th>
<th>Technologies</th>
<th>Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communities of interest</td>
<td>volunteers and amateurs</td>
<td>Coordinated interaction</td>
<td>self-convocation</td>
<td>blogs, wiki, file sharing</td>
<td>HERMEN</td>
</tr>
<tr>
<td>Communities of practice</td>
<td>technological, economical &amp; social areas operators</td>
<td>Coordinated interaction</td>
<td>self-convocation and ONG</td>
<td>blogs, wiki, free &amp; paid reservoirs and Web Portals</td>
<td>HEURIST &amp; HERMEN</td>
</tr>
<tr>
<td>Invisible colleges</td>
<td>professionals, technicians, specialists, scientists</td>
<td>not evident collaboration, but reciprocal influences</td>
<td>Academics and ONG</td>
<td>Local Networks (LNs), intranet: KMS, CMS, LMS, Web 2.0 growth</td>
<td>HEURIST &amp; HERMEN</td>
</tr>
<tr>
<td>Research &amp; Development teams</td>
<td>professionals, technicians, specialists, scientists</td>
<td>little coordinated interaction, primarily hierarchical structure</td>
<td>Academics and ONG</td>
<td>LNs, intranet: KMS, CMS, LMS, Web 2.0 growth</td>
<td>HEURIST</td>
</tr>
<tr>
<td>Groups of diffusion and communication</td>
<td>journalists, writers, artists</td>
<td>coordinated and cooperative interaction</td>
<td>enterprise, academics, ONG</td>
<td>blogs, wiki, file sharing and CMS</td>
<td>HERMEN</td>
</tr>
<tr>
<td>Communities of protagonists</td>
<td>special need people, war, abuse or disease victims, women, migrants, homosexuals</td>
<td>coordinated and cooperative interaction</td>
<td>self-convocation- ONG</td>
<td>blogs, wiki, file sharing and CMS</td>
<td>HERMEN</td>
</tr>
<tr>
<td>Communities of formation</td>
<td>teachers &amp; students, e-learning &amp; support of face-to-face training</td>
<td>little coordinated interaction, primarily hierarchical structure</td>
<td>enterprise, academics, ONG, public administration</td>
<td>LMS, Web 2.0 growth</td>
<td>HEURIST</td>
</tr>
</tbody>
</table>
2. Semantics in KOSs and language games

Knowledge representations incorporated in retrieval languages are often derived from assumptions originating from an empiricist view of knowledge, and from logical positivism (Svenonius, 2004). This is quite visible in the most frequently adopted approaches to meaning: operationalist or referential. And consequently in how KOSs — semantic tools providing meaning representation for a practical purpose — are conceived. Thesauri standards, for example, emphasize the need to include in a thesaurus only logically-based relationships among terms, which are considered permanent and valid in all contexts (a priori), such as genus-species.

Contemporary hermeneutics, instead, conceives knowledge as occurring always from within a situated perspective, a given horizon. Accordingly, meaning is not regarded as a fixed entity, but as socially and culturally constructed, open to interpretation and changeable according to different contexts (Hjørland, 2007).

The philosophy of the late Wittgenstein (1953) shares some similarities with parts of hermeneutics (above all Gadamer’s theories). Insisting on the fact that speaking language is a social action, he did not conceive the meaning of a word as an essence (i.e., a number of key conceptual features linked to a linguistic expression) or a referent, but rather as its use (as part of an activity) according to the rules of a given language game. Language is, thus, regarded as formed by an indefinite, however extensible variety of substructures (language games). Meaning is multidimensional, (to some extent) contextually determined and variable over time.

Together with hermeneutics, Wittgenstein’s theories, can be highly relevant to a series of issues in IS and KO (Andersen & Christensen, 1999; Blair, 1990; Hjørland, 1998). Above all, they can foster an idea of semantics in KOSs that, instead of searching an ‘out of context’ validity, attempts rather to represent (taking into account the actual needs of users) the complexity of knowledge and meaning as this occurs within different knowledge domains, and cultural and linguistic environments.

2.1 Possible applications to multilingual KOS development

In Wittgenstein’s view, concepts and meanings are created in the framework of social practices. Language games are, in fact, inscribed in forms of life, i.e. patterns of socially determined behaviours that encapsulate how we act and relate to the world as cultural communities and in which language arises and acquires meaning. Each culture is, thus, engaged in its own language games intertwined with their corresponding forms of life. In a way, expressing how words are used, and therefore their meaning, implies to depict a culture.

Hence, moving from a language setting to another could entail not only differences in the relations between concepts and their verbal expressions, but also at a conceptual level, since concepts and semantic structures are always developed according to the viewpoint of a given culture (or form of life).

In designing multilingual KOSs, such as thesauri, this should, therefore, be taken into consideration. Hudon (2001) has extensively explored this topic. As an example she provides the case of the French term éducation and of the English term education, which should not be regarded as exactly equivalent. The latter, in fact, has a broader meaning than the former, and is less clearly distinguished from the concept of teaching than éducation and enseignement are in French.
When comparing concepts/terms in different languages not only the semantics of the single terms should be considered, but more holistically also the way in which they interact with the semantic structure they directly take part in.

According to the semantic field theory of the structural linguist Trier (1931), a lexicon is, in fact, organized in clusters of semantically related and interdefined words. Similarly to Wittgenstein’s language games, this theory conceives meaning as arising from the wider context in which words are located: the semantics of any word results from the network of relationships among this word and others of the same field.

In this sense, each difference or modification occurring at the level of single terms can affect (and reorganize) the whole semantic field to which they belong. Therefore, interlinguistic equivalence should not be evaluated solely at a single term level. What matters is also how, moving from a language to another, the local semantic arrangement varies.

Overlooking the fact that diverse languages use differently terms which apparently express an identical concept could have consequences for the thesaurus semantic arrangement and lead to the establishment, for example, of controversial hierarchical structures, as shown below:

As stressed by Hudon (2001), French native speakers going through the French structure may doubt of its correctness. Besides its usefulness for meaning representation and navigation might be arguable, too. In fact, having (erroneously) established a full equivalence between education and éducation, this has been transferred to the following level of specificity.

This is what is likely to occur when the multilingual thesaurus is obtained by ‘translating’ in one or more languages an already existing monolingual (usually English) thesaurus, creating identical and symmetrical structures for all its linguistic versions. Nonetheless, while NTs of education are correctly placed, many of the Termes Spécifiques (TS) of éducation in reason of their semantics should instead be subordinated to other terms.
Education and éducation are used differently in their respective languages. In a sense, we may say that they are part of not (totally) equivalent language games. In the two languages the correspondent semantic field is differently carved.

To support effective multicultural indexing and retrieval, in developing the vocabulary and the semantic organization of the different linguistic versions of a KOS, the way in which each language is structured into its specific language games should be taken into consideration and (reasonably) represented, according to the real IR needs.

In order to reach such a purpose, the idea of a non-symmetrical multilingual thesaurus — i.e., consisting of independent terminological and semantic structures for each linguistic version (Hudon, 2001) — seems worth a further investigation. The notions of language games and semantic fields, together with hermeneutics, can contribute to reinforce the theoretical background according to which these new forms of multilingual thesaurus are designed. In any case, when structures are allowed to differ, they are likely to better reflect the conceptual and terminological settings with which potential users coming from different cultures are most familiar.

Conclusion

In this paper, we have tried to show how hermeneutics and other related theories could supply new insights into KO. Above all, they can provide a more realistic representation of the complexity of knowledge and language, which occur according to a variety of viewpoints, language games and forms of life. And consequently catalyze the creation of systems which incorporate this vision.

References


Abstract
Intertextual semantics is a semiotics-based approach to the design of communication artefacts primarily aimed at modeling XML structured documents. SKOS (Simple Knowledge Organization System) is a specification currently under development at the W3C that allows expressing various types of controlled vocabularies in XML. In this article, we show through an example how intertextual semantics could be applied to controlled vocabularies expressed in SKOS, and argue that it could facilitate the communication of meaning among the various persons who interact with a controlled vocabulary.

Introduction
Rebecca Green (Green 2001, 5) notes that one option for specifying the semantics of a relationship type in knowledge organization tools is to “simply [name] the relationship type that holds. This option makes the assumption that the user is familiar with the relationship type […] through his or her own personal experience and can access an understanding of the semantics of the experience on the basis of a natural language label.” The example that Green gives is the relationship type Bibliographic unit < is part of > Bibliographic unit. The natural language label chosen for the relationship type “is part of” is assumed to be sufficient to satisfactorily convey the semantics of the relationship type.

Green then argues (quite justly, in our opinion), that the seemingly quite different option that consists in specifying the semantics through formal language (rather than natural language) is, in reality, not very different from the natural-language label approach, because, ultimately, “the symbols of that formal language must be transformed into understanding on the basis of personal experience and probably using the medium of natural language.”

The idea that natural language might play an essential role in any approach to define the semantics of relationship types is very close to one of the hypotheses underlying intertextual semantics (IS), one of several semiotic-based approaches to the design of communication artefacts (De Souza 2005, Knuth 1984, Norman 2004). That hypothesis is that one of the most useful representations of a data structure populated with contents, for assessing—or even predicting—its usability and efficacy, is in the form of a natural language sentence (or “quasi-sentence”), or sequence of sentences.

Intertextual semantics
The goal of most semantic frameworks is to assign meaning to natural language, often through some sort of logic. Intertextual semantics pursues an entirely different goal. It assigns a meaning (semantics) to populated data structures, meaning which is expressed in natural language (NL). It involves the specification, by the creators of the data structure, of NL segments and of composition rules, associated with the various parts of the structure, allowing the meaning (in NL) of a given instance of the structure to be generated automatically and, for example, presented to a human user (author, reader, etc.). The objective of IS is to facilitate a common understanding of the instance among the various
human persons interacting with it throughout its entire lifecycle, including the persons who create the structure (the modelers), the persons who populate the instance (the authors), and the persons who consult it (the readers).

IS was introduced in (Marcoux 2006), where it was considered only from the perspective of modeler-author communication, and in the context of valid structured documents (e.g., XML). In (Marcoux & Rizkallah 2007b), it was applied to a more classical database-like structure, again with only the facilitation of modeler-author communication in mind.

The remainder of this section presents IS as introduced in (Marcoux 2006).

IS requires that modelers (XML document engineers, architects, etc.) document their models (DTDs, schemas, etc.) with artefacts allowing a “reference interpretation” to be generated automatically from any conforming document. That reference interpretation is in NL; it is, essentially, a character string making up a sentence, a paragraph, or a whole text. The underlying idea is that, with such an apparatus, the “intentions of the modelers” are easily communicated to human authors, right in the editing window, at the creation time of valid XML documents. This idea might be expressed bluntly as: “showing authors the immediate textual context in which their content is going to be interpreted, is the best way to get pretty much exactly what you want from them.”

As stressed above, IS uses NL as its semantic domain, not as its syntactic domain. The range of the semantic function is uninterpreted NL expressions, ready, though, for interpretation by humans belonging to some target community (whose definition likely involves the mastery of some given natural language).

The modeler-specified “artefacts” referred to earlier, consist of peritexts: one “text-before” and one “text-after” segment for each XML element-type in the model. The NL “meaning” of any conforming document is obtained by concatenating those segments to the actual textual contents of the XML elements in the document instance. The word “intertextual”—after “intertextuality,” coined in 1966 by Bulgarian philosopher Julia Kristeva—is used to reflect the idea that meaning is given to a piece of data by placing it in a network of texts or text segments.

Modeler- and author-contributed segments are assumed to be distinguishable from each other (for example, they could be of different colors). Modeler-contributed segments can contain some non-graphic characters (e.g., paragraph breaks) and the output of the semantic function as a whole can contain hyperlinks, in the form of URIs delimited by agreed-upon modeler-contributed markers, for example [square brackets]. Handling of XML attributes is discussed in (Marcoux & Rizkallah 2007a). The reader is referred to that paper and to (Marcoux 2006) for details and examples.

Related work
The idea of using text-related techniques to improve systems design, though not widespread, is not new: Smith (1994), for example, wrote that “[t]alk, theorized as conversation and analyzed as discourse, may provide the models of interaction that we need, in order to improve the design of hypertext systems and to extend the reach of its applications” (p. 281). Well-known examples of text-related techniques for systems development are Donald Knuth’s WEB system and Literate Programming in general (Knuth 1984), TEI’s ODD (One Document Does it all)—see for instance (Cover 2005)—, and SWEB (Sperberg-McQueen 1996). Using semiotics in general—not just text—in interface design has also been explored, for example by De Souza (2005). Personas (Pruit & Adlin 2006) and storytelling (Erickson 1996) also fall in that category of techniques.
Several formal approaches have been proposed in the past to provide semantics to structured documents. Renear, Dubin, and Sperberg-McQueen (2002) provide a historical background and a description of a specific project: BECHAMEL. Wrightson (2001, 2005) used situation semantics (Devlin 1991) to analyze, among other things, human legibility of XML well-formed documents. Another approach, with which IS shares many traits, is that of Sperberg-McQueen, Huittfeldt, and Renear (2000), where the authors develop a framework for structured-document semantics based on sentence skeletons and deictic expressions (roughly speaking, relative addresses). There are, however, important differences with IS.

SKOS and controlled vocabularies
SKOS (Simple Knowledge Organization System, Miles & Bechofer 2008) is a specification under development at World Wide Web Consortium (W3C), where it is part of the Semantic Web effort. It is “a common data model for sharing and linking knowledge organization systems via the Semantic Web.” The “knowledge organization systems” targeted by SKOS include thesauri, classification schemes, subject heading systems, subject indexes, and taxonomies. Elsewhere—for example, in (Miles & Pérez-Agüera 2007)—such knowledge organization tools are collectively referred to as controlled structured vocabularies or controlled vocabularies; in SKOS itself, they are referred to as concept schemes. We refer to them as controlled vocabularies (CVs).

SKOS is defined using two other languages of the W3C Semantic Web: RDFS (RDF Schema, where RDF stands for Resource Description Framework) and OWL (Web Ontology Language). One consequence of this is that SKOS CVs are actually RDF-compliant abstract objects, namely RDF graphs. For the purpose of encoding such graphs into concrete computer-manipulable objects, a number of notations (also called “serialization syntaxes”) can be used, through which the abstract objects are represented as character strings (or text files). Two such notations are RDF/XML and Turtle (Terse RDF Triple Language). RDF/XML is a W3C recommendation, whereas, at the time of this writing, Turtle is still a “team submission” at the W3C.

The RDF/XML notation expresses RDF graphs—and, thus, SKOS CVs—as XML documents. Even when a CV is expressed in Turtle, it is a trivial matter to convert it to RDF/XML. Without loss of generality, then, we will assume here that SKOS CVs are always expressed in RDF/XML. This allows us to discuss SKOS CVs within the IS framework, as defined in (Marcoux 2006) for XML documents. We will further assume that any given SKOS CV corresponds to a single XML document. This is always possible with RDF/XML, though not mandatory: a CV could theoretically be spread over any number of XML documents.¹

¹. A whole CV as a single document would likely be large. In the IS framework, the “meaning” of that document could be a very large NL segment. Should a user choose to access the whole document in “IS view,” navigating through this large NL segment might be somewhat difficult. Concrete environments could alleviate this problem by providing the possibility—mentioned in (Marcoux, 2006)—of expanding only selected fragments of the document into their IS-view, and that of expanding and collapsing on demand the hierarchical structure of the document (a feature present in most extant XML authoring and viewing tools). Internal navigation possibilities could also be embedded directly in the output of the semantic function, in the form of internal hyperlinks.
Controlled vocabularies and intertextual semantics

A person can interact with a CV in a variety of ways. For example, an information professional may navigate through it in order to add terms and relationship instances, whereas a user might browse it to locate resources relevant to an information need. Some interactions actually involve not the CV as a whole, but only part of it; for example, when a user browses the various terms that have been assigned to a resource.

Although it is possible to formulate an IS for a CV “absolutely,” without reference to any particular interaction type, it is simpler if we have a specific interaction type in mind. Here, we will target the situation in which an information professional navigates through the CV to add terms and relationship instances.

We illustrate how IS can be used with CVs through an example, and subsequent discussion. The example is adapted from (Mikhalenko 2005) and corresponds to a tiny portion of the *Alexandria Digital Library Feature Type Thesaurus* (http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/ver070302/index.htm). First, we present the raw SKOS document (in RDF/XML), then one possible IS view of that document (the precise IS mechanism has been somewhat loosened in the example, for the sake of clarity). The IS specification itself can be easily inferred. See (Marcoux & Rizkallah 2007a) for examples of IS specifications.

Example — Raw SKOS document:

```xml
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:skos="http://www.w3.org/2004/02/skos/core#">
  <skos:Concept rdf:about="http://www.my.com/#canals">
    <skos:definition>Manmade waterway used by watercraft or for drainage, irrigation, or water power</skos:definition>
    <skos:scopeNote>A feature type category for places such as the Erie Canal</skos:scopeNote>
    <skos:prefLabel>canals</skos:prefLabel>
    <skos:altLabel>ditches</skos:altLabel>
    <skos:altLabel>drainage canals</skos:altLabel>
  </skos:Concept>
  <skos:Concept rdf:about="http://www.my.com/#hydrographic%20structures">
    <skos:prefLabel>hydrographic structures</skos:prefLabel>
  </skos:Concept>
</rdf:RDF>
```

2. In practice, the various interaction types would take place on different documents, derived automatically from one another, and/or from the CV itself. Each of those documents could have its own model and IS. The important thing is that the different ISs be pairwise consistent and consistent with the IS of the CV.
Example — IS view of the SKOS document:

[... Introductory section for the whole CV: background, purpose, scope, etc. (omitted) ...]

Section for concept with formal identifier: http://www.my.com/#canals

This concept can be defined as *Manmade waterway used by watercraft or for drainage, irrigation, or water power.*

It can be used as *A feature type category for places such as the Erie Canal.*

The official accepted word or expression for referring to this concept is *canals.*

Another word or expression commonly used to refer to this concept is *ditches.*

Another word or expression commonly used to refer to this concept is *drainage canals.*

*Canals* are special cases of *hydrographic structures.*

End of section

Section for concept with formal identifier: http://www.my.com/#hydrographic%20structures

The official accepted word or expression for referring to this concept is *hydrographic structures.*

End of section

Discussion

First, note that the contents of the IS goes beyond the inferences that SKOS strictly allows us to draw from the CV. For instance, the interpretation given in the IS to the skos:broader relation type assumes that it is a relation of type “is-a” (a canal is a hydrographic structure), which is *not* implied by the definition of skos:broader. That illustrates exactly the kind of benefits that specifying the IS of a CV might bring: if it is the case that the actual semantics of a relationship is richer (or more refined) than what is inherent in SKOS (which, by the way, is essentially the same as what is found in CV construction standards), then there ought to be a way to convey that extra semantics to the users.

We mentioned earlier that IS is about facilitation of the communication between various people interacting with a data structure. In our case, the communication is between a person we call the modeler of the CV and a person we call an author of the CV. The modeler is the person who determines the structure of the CV, i.e., what semantic relations it includes, what their properties are (transitive, antisymmetric, etc.), what the types of label and of documentation elements are, etc. Then, the modeler maps those elements to a specific framework, for example SKOS or various standards for the construction of CVs. *It is at that time that the modeler should produce the IS specification for the CV,* to document the fact that some extra semantics (if any) is available, and document it in a way that can be delivered to future users of the CV. Among which are the authors of the CV.

An author of the CV is a person who enters terms and relationship instances in the CV (with or without automated means). For the CV to be as consistent, complete, and useful as possible, it is important for the authors to understand exactly the decisions taken by the modeler. In our example, the author is the person for whom the IS of the CV is intended. The IS specification prepared earlier by the modeler is mixed with the actual contents of the CV, thus forming the IS view presented to the author (the author-contributed segments are shown in italics in our example). If the modeler has done his job correctly, this IS view conveys to the author all the semantics allowed by the modeler’s decisions, not just the semantics strictly allowed by the standard chosen to construct the model.
The IS specification used for our example was of course not produced by the modeler, it was produced by us, the authors of this paper. We actually had to guess whether there was any extra semantics the modeler intended to put in his model, by observing the excerpt we had under our eyes. For example, in the excerpt observed, it appeared that the relationship coded as skos:broader was actually a “is-a” relationship. But the only way to be sure would be to hear it from the modeler. If producing an IS specification were part of standard CV construction methods, there would have been no need for guesswork, neither from us, nor from users who might want to know exactly what (or, we could say, how much) they can infer from the CV.

SKOS proposes a standard structure for CVs, i.e., a defined set of semantic relations, and of label and documentation elements. As mentioned earlier, this structure is essentially the same as that found in CV construction standards. However, SKOS also allows the definition of customized structures, through extensions (expressed in RDF, RDFS, and/or OWL). If the modeler makes use of such extensions, then the IS specification is even more important, because there is no “standard” interpretation to which the features obtained by extension can be linked. We are 100% dependent on the modeler’s explanations to understand those features of the CV.

Note that the linguistic formulations found in the IS specification are often awkward: this is because models are not written with IS in mind. If producing an IS specification were part of standard CV construction methods, we would probably observe that formulations in the IS would be simpler.

In first analysis, it might be tempting to “shove” to the user-interface the task of conveying to the users whatever extra semantics there might be in the structure of a CV, thereby relieving the modeler of the burden of producing an IS specification. We think this is actually begging the question, because, then, the issue becomes that of conveying the intentions of the modeler to the interface designer. For this, something essentially equivalent to an IS specification seems unavoidable.

**Conclusion**

In this article, we showed through an example how IS could be applied to CVs expressed in SKOS, and argued that it could facilitate the communication of meaning among the various persons who interact with a CV. Research must be pursued, both theoretically and practically, to establish more precisely the actual benefits that might result from implementing the ideas expressed here.

It is not clear exactly at the time of this writing how IS would be best integrated into SKOS, because not all extension mechanisms of SKOS have been defined yet. An extension-by-refinement mechanism is described in (Miles & Pérez-Agüera 2007), and will probably be included in the final version of SKOS. One thing we can say is that the IS peritexts are not refinements of any of the existing documentation properties of SKOS, and should therefore not be defined as such, otherwise, it could lead to incorrect inferences being derivable from the CV. A more promising avenue seems to be the use of XML elements and/or attributes of domain-specific XML namespaces.

**References**


Language Related Problems in the Construction of Faceted Terminologies and their Automatic Management

Abstract
The paper describes current work on the generation of a thesaurus format from the schedules of the Bliss Bibliographic Classification 2nd edition (BC2). The practical problems that occur in moving from a concept based approach to a terminological approach cluster around issues of vocabulary control that are not fully addressed in a systematic structure. These difficulties can be exacerbated within domains in the humanities because large numbers of culture specific terms may need to be accommodated in any thesaurus. The ways in which these problems can be resolved within the context of a semi-automated approach to the thesaurus generation have consequences for the management of classification data in the source vocabulary. The way in which the vocabulary is marked up for the purpose of machine manipulation is described, and some of the implications for editorial policy are discussed and examples given. The value of the classification notation as a language independent representation and mapping tool should not be sacrificed in such an exercise.

Facet analysis as a general tool for vocabulary construction
The value of using a faceted classification as a basis for a thesaurus has long been acknowledged, primarily in the work of Jean Aitchison (1986) who pioneered the methodology and built several subject specific thesauri (Aitchison et al., 1992) using BC2 schedules as a basis.¹ The role that can be played by facet analysis in the construction of thesauri is now formally acknowledged in recent revisions of the thesaurus standards. For example, in the new British Standard, BS8723 the value of facet analysis as a general methodology is explicitly stated:

Facet analysis is useful in generating hierarchies that conform to the rules for hierarchical relationships ... because these relationships are valid only for terms belonging to the same general category. ... The choice of facets may vary ... but ... it is usual to use fundamental categories such as objects, materials, agents, actions, places, times, etc. These fundamental facets may be analysed into subfacets where it is helpful to do so ... (BS8723-2:2005, p. 31)

This is followed by further discussion of the detailed analysis, and accompanied by figures showing the faceted display with its subfacets (or arrays) and node labels (principles of division). Facet analysis is considered at various other places in the standard (BS8723-2:2005, p. 32, 36, 37, 38, 39, 41), usually as part of a process of thesaurus construction that follows the Aitchison model. The current American standard (ANSI/NISO Z39.19-2005) also mentions facet analysis for the first time, although it is more cautious in its approach, stating only that facets may be useful (ANSI/NISO, 2005, 14–15). Although facet analysis is understood differently in the US (ANSI/NISO 2005, p. 14), it is clear that, in practice, facet analysis is applied in much the same way as in the UK, and the Art & Architecture Thesaurus is used to provide examples. Facet analysis is mentioned at several points in the text, including the observation that it is typically used for ‘large controlled vocabularies covering a broad domain or discipline with complex relationships among terms’ (ANSI/NISO 2005, p. 141).

The faceted classification, because of the rigorous analytical principles used in its construction, makes explicit most of the relationships employed in the standard thesaurus

¹. These included not only the international affairs thesaurus cited, but also the Department of Health and Social Services thesaurus, which was based on the Health Sciences Class H of BC2.
format. Because individual concepts will have a particular facet ‘status’ (i.e. they will be named as property, process, agent, etc.), these relationships usually can be inferred from the systematic display. The hierarchical or paradigmatic relationships between concepts in the same facet correspond to the BT/NT relationships of the thesaurus; additionally, relationships between concepts in different facets can be assumed to be syntagmatic in nature, and to fall within the scope of the associative, or related, term relationship. Collocation of synonyms and synonymous phrases within a class heading, and represented by the notation, can be regarded as a form of vocabulary control comparable to equivalence relationships in the thesaurus.

A particular constraint on the current work with the Bliss Bibliographic Classification 2nd edition (BC2) is the desire to produce an integrated classification, alphabetical index, and thesaurus. This demands that the three aspects of the faceted terminology are coherent, and that the characteristics of one cannot be compromised to achieve the required results in another. We believe that the faceted methodology is sufficiently robust to support all three formats, but that factors associated with language control have not been adequately addressed in constructing the classification, and need now to be incorporated into the process (Aitchison, 2004).

Automatic generation of a thesaurus from a classification

In contrast to earlier faceted thesauri which have been built manually, over recent months we have been developing a semi-automatic means of generating the thesaurus from the schedules of BC2 (Broughton, 2008). BC2 already has a suite of programmes which take a raw data file and convert this to a properly formatted classification schedule, and which also generate the alphabetical subject index. The programmes manage the physical display of the classification from information about the hierarchical level and sequence of classes contained in the source file, and also recognize ‘non-classes’ such as facet and array labels and handle these accordingly. The index programmes manipulate the class headings to provide suitable index entries, and perform a useful function in identifying duplication and errors. The information necessary for these operations is inserted into the source files using a basic kind of mark-up language or encoding.

Here, notational codes and symbols, indent codes, punctuation symbols and other marks are instructions to the software. The file consists of both classes and non-classes (such as principles of division, or node labels) known collectively as items. The class mark notation acts as a ‘control number’ for each class, while the @ symbol denotes a non-class. Each item has a line number in the file, and there is also an element of mark-up which indicates the position in the hierarchy i.e. level 1, level 2, etc. These are known as indent codes, and they serve, in the first instance, to define the indentation in the physical display of schedules. The sample source file for Chemistry demonstrates the indent codes (06, 07,08, etc.). Letter codes and punctuation symbols suffixed to the items direct the appropriate programs to include, exclude, modify or invert the items in the alphabetical index. By this means, features of the finished classification, such as the indentation, fonts, and carry-over hierarchies, and the selection, qualification and inversion of alphabetical index terms, are achieved automatically. This works extremely well for the classification physical display and tolerably well for the alphabetical index, despite the need for some manual editing of the latter.
Extending the software to accommodate the thesaural relationships also works well at a structural level. The software is able to infer parent, child and sibling relationships, and hence broader and narrower terms (and some related terms), from the relative positioning of classes (determined by the indent codes and notational sequence) and its ‘understanding’ of the application of the organizational principles of the faceted scheme.

The software does this by taking information from the raw source file and converting it into a database. This data can then be manipulated by the other programs to produce intermediate files for manual editing, draft, layout and print files for the schedule, index and thesaurus. The database contains both ‘abstract data’ i.e. conceptual information relating to the nature of a given class, and ‘concrete data’ which determines how the information is to be formatted in any particular context. The concrete data is not unlike the code used in a word processing package to manage things such as the indentation, spacing, fonts and so on. The abstract data is more sophisticated than this, and allows the software to calculate the relationship of one class to another. To do this the software must have certain structural and other classificatory attributes of each item clearly defined. Classes and non-classes (such as facet and array indicators, or node labels) together constitute the items in the file. For each item it is necessary to know:

- its position in the hierarchy of items (shown by the indent code)
- its classmark, if it has one (the notational code)
- its names (terms used in the class descriptions)
- its cross-references, if it has any
- its importance indicator (a device to ensure appropriate column/page breaks), if it has one.
It is possible to calculate the hierarchical position of an item from its level (which may be indicated by coding or by regulated tabulation) and the hierarchic order (this being controlled by line numbers in the file and also by the notation). Because the general scheduling rules for a faceted scheme control the organization of classes into arrays and facets, the program knows that there is an invariable sequence of:

- a class
- its first offspring
- if no offspring, its next sibling
- if no more siblings, its parent’s next sibling
- if no more parent’s siblings, its grandparent’s next sibling
- and so on, ad infinitum.

This provides the hierarchic order, and given that knowledge, and the level of each class, it is possible to infer the parent, previous sibling, next sibling, or offspring of an item, and to ascertain the relationship between any given pair of classes. Consequently, it is also possible to infer many of the thesaural cross-references from this information, where these are supported by the facet structure as discussed above. Because the program can infer an upward reference, i.e. to a class’s immediate parent (or if the immediate parent is a non-class, to the grandparent), broader terms can be identified. Whenever an upward reference is inferred, the reciprocal downward reference is also inferred. Although upward and downward and BT and NT are not absolutely synonymous, they usually are, and this is sufficient for the program to infer these relationships unless otherwise instructed.

Vocabulary control in class headings

However, while the theory and its software application stand up well in terms of the structural and navigational aspects of the thesaurus, problems are encountered when dealing with the linguistic features that are essential to a successful thesaurus. This is hardly surprising, since in a classification the notation is used to represent the class and there is no need to consider or control the linguistic form of the class description, or to nominate preferred terms. A detailed analysis of examples where the concept or class heading could not be satisfactorily used as an entry term for the thesaurus allowed us to identify recurring patterns in the classification, and to begin to develop an editorial policy for the integrated tool.

The majority of these situations arise from BC2’s dependence on the notation to carry the semantic burden, with consequent shortcomings in the class description. Some typical examples include:

- incomplete terms where meaning is taken from the context i.e. ‘with dogs’ as a sub-class of the heading ‘police work with animals’, itself a sub-class of ‘agents of police work’ (correct term should be ‘police dogs’)
- the use of terms that are really combinations, but where only the ‘added’ concept is expressed e.g. ‘prices’ as a sub-class of ‘housing market’ (correct term should be ‘house prices’)
- failure to distinguish between synonym sets and sibling sets in a class heading.

The latter is potentially a more serious problem since it obscures the distinction between equivalence and associative relationships.
Managing equivalence relationships
Equivalence relationships are implicit in the faceted classification, and some elements of the schedule display in BC2 do serve to identify them. There are, for example, numerous examples of where the conceptual structure of the classification collocates synonyms (equivalent terms) within the same class heading or caption. In the examples in the following section of schedule, it is clear that an equivalence relationship is intended between the pairs of terms ‘musicals’ and ‘musical plays’, and ‘cello’ and ‘violoncello’, and so we may deduce appropriate UF and USE references:

<table>
<thead>
<tr>
<th>WWF O</th>
<th>Opera</th>
</tr>
</thead>
<tbody>
<tr>
<td>WWF R</td>
<td>Operetta</td>
</tr>
<tr>
<td>T</td>
<td>Musical plays, musicals</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Musical plays</th>
<th>UF Musical plays</th>
</tr>
</thead>
<tbody>
<tr>
<td>Musicals</td>
<td>USE Musical plays</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>WWQ Z</th>
<th>Bowed instruments</th>
</tr>
</thead>
<tbody>
<tr>
<td>WWS B</td>
<td>Cello, violoncello</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cello</th>
<th>UF Violoncello</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violoncello</td>
<td>USE Cello</td>
</tr>
</tbody>
</table>

For the most part the BC2 schedules observe these stylistic conventions for synonyms. However this is not followed through consistently and there are cases where the first term is not the most natural choice, and other cases where synonyms are omitted. The software does not recognize any priority where a class has several names, and unless other restrictions are applied, will treat all names of a class as equal. Similarly, although classes can be identified by either classmarks or names, it is implied that searching by name may be less effective than searching by classmark.

A further problem occurs when several concepts are grouped together simply for convenience, or because there is no easily named containing class, as in this example where the comma is used to separate coordinate members of the same class. At present there is no way to distinguish this from synonym groups.

| WWG F | Liturgical music |
| WWG Q | Motets, anthems, hymns, etc. |

Compound terms and semantic factoring
Another feature of the faceted classification (although this is less marked in BC2) is the limited size of the vocabulary in comparison with the thesaurus, and the stress on conceptual status. The implication is that compound terms (whether semantically or syntactically compound) are less likely to appear in the vocabulary of the classification, whereas they must be represented in the thesaurus. This is a particular problem in the humanities, where cultural differences tend to generate unique sets of terms when the primary facet is varied. This difficulty had already been encountered in the management of the religion vocabulary in the UDC revision of the Religion (Broughton, 2000) and in the FATKS project (Broughton & Slavic, 2007), where UDC and BC2 were used as a source of ter-
minology. The question of factoring (or splitting) which is considered in considerable detail in the thesaurus standards tends largely to be by-passed in the faceted classification. Semantic factoring is discouraged by the thesaurus standards, but may be normal classificatory practice. Similarly, the incorporation of large numbers of unique terms in humanities subjects is alien to the understanding of a faceted classification, but can hardly be ignored when a thesaurus is under construction. The music class of BC2, which was the pilot class for the thesaurus project, was initially devoid of any terminology other than that of the western classical tradition; expansion of the vocabulary with the thesaurus in mind disclosed the existence of very large numbers of terms peculiar to the music of individual cultures and movements as revealed in folk and world music, as well as in the many manifestations of popular music.

A faceted classification would tend to express such terms conceptually, representing them as the sum of constituent classes, and creating a notational code which is the equivalent of the compound term. Such a code, where the notation is expressive and its integrity is maintained, can support cross searching for the individual elements of the compound. For example, duduk is the name of a folk oboe in the Armenian tradition. Such a term might not be entered in a faceted classification, but left to the indexer to construct as a compound of ‘folk music — Armenia — oboes’. Although this would contravene good practice from a thesaurus standpoint, it has the advantage that the document would be retrieved in a search for ‘folk — oboes’, or ‘Armenia — oboes’, despite the fact that a search for ‘duduk’ would not be successful.

Related problems of data management occur when trying to use a faceted classification semi-automatically via a database, as was the case in the FATKS project. The existence of these culture dependent terms causes difficulties for cross-searching, but here the notation has a role to play in representing the conceptual structure of the terminology and supporting searching, even when it is hidden behind the interface and not visible to the end user. The database created to hold the classification data for FATKS accommodated an encoding system similar to that of the BC2 source files, but more extensive in that it also denoted the facet status of concepts as well as their hierarchical and linear positions. Here, the notation can be used as a control to manage the cross-searching and to sustain the conceptual structure, as in the example above. Because the facet membership of each concept is explicit, encoding of this type could enable the semi-automatic identification of syntagmatic terms in actual documents in a way that the BC2 encoding cannot.

**Conclusion**

The semi-automatic management of faceted terminologies and the generation of a variety of distinct but interdependent and coherent formats from the source vocabulary is shown to be viable, particularly in terms of the structural relationships. In BC2 encoding of the source data supports this operation and allows the inference of most of the thesaural relationships; encoding of the more elaborate kind used in the FATKS project (where categorical status was included) would permit inference of syntagmatic as well as paradigmatic relationships.

Because vocabulary control has not been a requirement in the construction of classification schemes, some intellectual intervention is necessary to correct existing anomalies in the BC2 terminologies, but a revision of the editorial rules would facilitate better automatic management. Potential conflicts exist between the classification and thesaurus formats, particularly in the approach to factoring, and where culture specific terminology
increases the size of the vocabulary. It is important to recognize that both the thesaurus approach, with stress on inclusion of distinct terms, and the classificatory approach, with its representation of the conceptual makeup of a term, offer different advantages in terms of searching and retrieval. The style of BC2 with an extended terminology organized by strict facet principles may offer a combination of the two, but much work needs to be done on establishing rules for formatting of class headings and the control of vocabulary before this can be managed completely automatically.

References
Broughton, V. (2008). A faceted classification as the basis of a faceted terminology. Axiomathes 18(2) July. Published online at: www.springerlink.com/index/6mm3r57j1r44k5u5.pdf.
I. C. McIlwaine — University College London
Nancy J. Williamson — University of Toronto

Medicine and the UDC
The Process of Restructuring

Abstract
This project to reconstruct and update the UDC Medicine Class (61) is an experiment in the ongoing development and revision of the Universal Decimal Classification as a whole. As part of that process, in 1988 the UDC Management Board recommended the establishment of a limited life Taskforce on UDC System Development to advise on the future development of UDC. That Taskforce recommended that a study be carried out to determine the feasibility of converting the system into a fully faceted classification. It was decided that the best way to accomplish this would be to convert one class. Medicine (Class 61) was chosen for the experiment for two reasons. First, this class was a part of the system that most out of date and greatly in need of revision, and secondly, it presented an opportunity to test an approach to the subject matter which would be in be in keeping with modern methods of the study and practice of medicine. This project is a direct response to that recommendation. Phase 1 of the project is now complete and Phase 2 is underway. This paper describes Phase 1 and its findings, identifies problems still to be addressed and sets out the methodology for Phase 2.

Introduction
Over its long history, the UDC has undergone many changes in its management and development. Its foundations are in Dewey Decimal Classification edition 5, but it has been developed independently from Dewey and with different objectives. It is a bibliographic classification primarily designed to organize bibliographic listings as opposed to organizing books on shelves. UDC inherits partial enumeration from Dewey but with its auxiliary tables and symbols “it is a synthetic classification, which means that the enumerated classes are the building blocks whereby compound and complex classes may be denoted by various synthetic notational devices” (McIlwaine 2007, 9). UDC was an inspiration for the development of Ranganathan’s Colon Classification (Ranganathan 1933). “A truly faceted classification does not permit the listing in schedules of any combination of terms and states each concept in simple terms. While it is not fully faceted, in that sense, the principles of facet analysis are inherent in the structure of UDC” (McIlwaine 2007, 33). For it to become so, it would need to completely shed its historical connections with Dewey (Dewey 1894). One UDC class where this has already been achieved is the recently approved Class 2 Religion. Theology. This project is a further attempt to move in that direction.

Support for this project is also found in the work of the Classification Research Group (CRG) in the UK. Following from Ranganathan, the CRG refined his principles and applied them in a number of special subject schemes and in the development of the Bliss Bibliographic Classification (BC2). That system is fully faceted and applies sound principles of facet analysis. While BC2 is still in progress, the schedule for Class H: Anthropology, Human Biology, Health Sciences has been published and is being used in this project.

Phase 1
Phase 1 began with the establishment of the following goals: to address the question of the feasibility of converting UDC to a fully faceted system; to provide an up-to-date UDC class for medical sciences; to create a table that can be used in conjunction with the existing UDC; to achieve a display of topics commensurate with the modern approach to
the study and practice of medicine; and to provide a table with depth of analysis commensurate with the recently published Standard Edition of UDC (2005–2006).

In support of these goals, the human biology and health sciences section of BC2 Class H (Mills, and Broughton 1981) was chosen as the basis for the framework of the proposed UDC Class 61 table. Its suitability for the task was judged to be its sound principles of facet analysis and its emphasis on the grouping of the subtopics of medicine by systems of the body. This approach is currently used in the study and practice of modern medicine. Because of the age of Class H (1981), it was recognized that additional support would be needed from such as tools as Medical Subject Headings (MeSH) and the ICD-10: International Statistical Classification of Diseases and Related Health Problems.

To establish the framework, the first step was to work with Class H schedule and to convert it to a UDC-like system. The Class H schedule was scanned and converted to machine readable form. The notation was stripped out and the data were divided into subclasses (e.g. curative medicine, clinical medicine, diseases and pathology, cardiovascular system, etc). Figure 1 below illustrates the BC2 format. The facets are shown in parentheses.

In Figure 2 below, the BC2 notation has been stripped out and replaced with UDC numbers. The topics have been arranged and subarranged. Built topics appear as example(s) of combination under a main caption. The facets are worked into the examples as parts of the built numbers and the remainder is built up from other subclasses in the 61 class, and where appropriate, from other parts of the UDC classification.

Figure 2 illustrates facets in four ways — in the captions by such phrases as “617.231 – by manifestation” and “617.232 – by cause”; as class numbers following the colon and taken from 615 Diseases and Pathology e.g. 617.231:615.2 – Neoplasms (malignant); in the numbers (e.g. -044.88) taken from the common auxiliary tables, and from the special auxiliary table of types of diseases in 615 (e.g. -216.12 oedema; -216.22 rupture and -216.42 obstruction). In the built numbers, the last number represents the facet. The number 579.862 represents the name of the virus taken from the 570s in UDC and 618 … indicates an undeveloped section. In all, 14 subclasses were developed in this manner. All have been published as proposals in the annual publications Extensions and Corrections to the UDC, between 1993 and 2007.

Findings from phase 1
The resulting table is not a final scheme ready for use. Rather, it is basically Class H interpreted in the light of UDC and designed to act as a framework for the final product. The result is a flexible and workable system, which will accept new built numbers as needed and should not require frequent major revisions. The use of facets works particularly well in the tables on the various body systems where many of the diseases and other medical problems are contained under the label example(s) of combination. When new topics emerge, additional examples can be created easily. The tables are rich in terminology and this has had some affect on UDC as a whole. In particular, Phase 1 was instrumental in identifying some of the deficiencies in the tables of common auxiliaries and the work on Class 61 has provided considerable input to the development of tables 1k -02 Common auxiliaries of properties and -04 Common auxiliaries of relations, processes and operations which were created during the project. Not unexpectedly, the notation needs further attention and it was known from the outset that there would be major updating of data at phase 2. Also, clarification of the relationships between the proposed 61 tables and the rest of UDC is
needed. However, the project to this point has been relatively successful. Now there is a
workable base from which to complete the work on the proposed Class 61, but there is
still much to be done.

**Phase 2**
The objective of Phase 2 is to bring the proposed Class 61 to a workable conclusion. Ini-
tially, two questions need to be addressed: What exactly should be the content of the final
version of the new medical class? What implications does content have for the rest of
UDC? The answers to these questions are fundamental to the completion of the final ver-
sion. The circumstances under which UDC 61 and BC2 Class H were created differ. UDC
class 61 is a part of the whole of UDC, whereas in BC2 each class has been developed to
stand alone. Thus Class H contains topics that make it “whole”, whereas in UDC some of
these topics are dealt with in other classes, so there is an inevitable “knock-on” effect. For
example, BC2 Class H has a substantial section on biochemistry. Should this be retained
under medicine? Or should it, more logically, be covered in 57 with the rest of biochem-
istry? Biochemistry in UDC is another class that is sadly out of date, which means a major
revision. In Class 61, there are some topics that have been split. For instance, the embry-
ology of plants is in 581.3, of animals in 591.3, while human embryology is 611.013. On
this, UDC and BC2 agree. Both put human embryology with medicine. Should the split
be continued in the proposed 61, or should all embryology be together in 57 and the class
number pulled forward into built numbers, as needed. In a faceted classification, it would
seem that there might be some logic in the latter position. Numerous other cases of such
problems exist and will need to be addressed in Phase 2. A brief survey will be conducted
to determine how best to handle them. There will be a need for matching process between
the proposed Class 61 and the whole of UDC to answer such questions as: What topics in
the existing UDC 61 are missing from the new version? What topics related to medicine in
the proposed new 61 are located elsewhere in UDC? What topics in other classes in UDC

---

**Figure 1. BC2 Format**

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUH</td>
<td>Heart, cardiology</td>
</tr>
<tr>
<td></td>
<td>(Disorders by manifestation)</td>
</tr>
<tr>
<td></td>
<td>(Dysfunction)</td>
</tr>
<tr>
<td></td>
<td>*For arrhythmia see disorders of</td>
</tr>
<tr>
<td></td>
<td>of physiological processes HUH OJ</td>
</tr>
<tr>
<td></td>
<td>(Enlargement)</td>
</tr>
<tr>
<td>HUH JK</td>
<td>Heart enlargement</td>
</tr>
<tr>
<td>HUH JL</td>
<td>Cardiac oedema</td>
</tr>
<tr>
<td>HUH JO</td>
<td>Heart rupture</td>
</tr>
<tr>
<td>HUH KA</td>
<td>Congestive heart failure</td>
</tr>
<tr>
<td></td>
<td>(Paroxysmal dysporoea) see Respiratory disorders HWE OXX</td>
</tr>
<tr>
<td></td>
<td>(Neoplasms)</td>
</tr>
<tr>
<td>HUH ME</td>
<td>Carcinoid heart disease</td>
</tr>
</tbody>
</table>
Figure 2. UDC Format

617.231 Heart disorders by manifestation

→ 617.243:616.72 Congenital disorders of the myocardium

Example(s) of combination:
617.231:615.2-021.473 Neoplasms (malignant). Carcinoid heart disease
617.231-044.88 Enlargement of the heart
617.231-214.11 Heart dysfunction
→ 617.233:617.211 Arrhythmia
617.231-216.12 Cardiac oedema
617.231-216.22 Heart rupture
617.231-216.42 Obstruction.
Including: Congestive heart failure

For Paroxysmal dysporoea see 618... Respiratory disorders

617.232 Heart disorders by cause

Example(s) of combination:
617.232:615.3 Heart infections
617.232:615.4:579.862 Streptococcus haemolytica. Rheumatic heart disease

need to be revised or added to? This will require considerable cross checking between Class H, the two versions of class 61 and the whole of UDC. Most of the problems of this kind will occur in the subclasses on anatomy and physiology, health science (general), preventive medicine, curative medicine and clinical medicine. This will require the removal of some topics from medicine and major revisions in other areas of UDC.

There are several kinds of updating needed — internal updating of the new tables; updating with respect to the Standard Edition; updating of the diseases; and making sure that all topics in the present Class 61 are present or accommodated for in the proposed tables. Internal updating of the proposed tables is required because the various subclasses were created in serial order. Once a subclass was published we did not return to revise it, as ongoing revisions were made in UDC. Some class and table numbers chosen early in the project were superseded by ongoing revisions to the existing tables and the development of new common auxiliary tables. Hence there is an unevenness of application and numerous blanks exist where appropriate class numbers were not available; for example the two important common auxiliary tables, -02 and -04, that were not available at the beginning of the project. There are numerous blanks and incomplete cross references to other parts of the proposed 61 which were developed later in the project. While the proposals were being developed, the rest of UDC was in continuous revision. Moreover, the topics contained in the proposal are based largely on BC2. Albeit, BC2 has more detail than the current 61, but are all the topics in the current 61 covered? What are the differences in terminology? In summary, the proposed medicine class needs to be brought into line with the latest Standard Edition.

The updating problem is even more complex, in that the present Class 61 is very outdated — a reason for selecting it for this project. But BC2 Class H was published in 1981. Much has happened in medicine and the identification and treatment of diseases in the intervening 27 years. One of the most important tasks in Phase 2 will be the updating of proposed subclass 615 Diseases and Pathology. Subclass 615 lists the major diseases and is
fundamental to the development of the subclasses on the various body systems (616/619). General diseases listed here are taken as they apply to diseases and illnesses identified as attacking different parts of the body. Cancer (neoplasms), for example, has a basic class number 615.2 and this number is used under each body system together with the name of the type of cancer appropriate to that part of the body. Neoplasms of the heart are in “617.231:615.2-021.473. Neoplasms (malignant) Carcinoid heart disease.” Lung Cancer appears as “618.531:615.2 Lung neoplasms. Including: Bronchiolar carcinoma. Bronchogenic carcinoma. Pancoast syndrome”. Other tools are needed to aid in this segment of the revision. The World Health Organization’s *International Statistical Classification of Diseases and Related Health Problems*, supplemented by the National Library of Medicine’s *Medical Subject Headings* will be used. The WHO publication will be most helpful as it is subarranged by systems of the body and is in classified order. Also particularly useful will be the sections of the print edition of MeSH which list subject headings added annually to the system from 1981 to 2003. Both are up-to-date and available online.

Another aspect of the revision is the use of special auxiliary tables. In the present 61 there are numerous special auxiliary tables. There are only two special auxiliary tables in the proposal for diseases in 615 and for regions and parts of the body in 616. Should there be more? Are there other subtopics which should be dealt with in this way (e.g. medical equipment?) Or should all equipment be listed in 68 under manufactures and coloned as appropriate?

Finally, there are concerns about the notation which must be addressed. At the very least, the notation needs to be adjusted. At the outset, it was impossible to determine precisely how much space would be needed to handle each of the subclasses. They are unequal in content and, not unexpectedly, there is some crowding at the end of the class. While it can be anticipated that some sections may be moved elsewhere freeing some space, it may not be sufficient and there is a need to allow for the future. As far as the class as a whole is concerned, much additional growth may be due to the addition of new diseases. Some of this will be absorbed in the examples and may not require new class numbers but this cannot be counted on as a certainty. One suggestion is that medicine be moved to 4 which is currently vacant, thus providing a 100 number base instead of a base of 10. Long class numbers are a problem in some cases and some people feel that a full 100 number base would contribute to shorter class numbers. This may be partly so, but the depth of analysis also has a bearing on the length of class numbers. While the intention is to equate Class 61 with the UDC Standard Edition, in its final form, Class 61 will still be very detailed.

**Conclusion**

Phase 1 appears to have resulted in a usable framework based on sound principles which should make the next steps easier to accomplish. While there is still much detailed work to be done, the move toward the final result should be accomplished much more quickly than Phase 1, because there is now a basis for it. Once the work is completed, it still has to be submitted to a panel of experts for review and analysis.

**References**


Maltby, A. 1975. Sayers’ manual of classification for librarians. 5th ed. Andre Deutsch., p. 34.


Cultural Markers and Localising the MIC Site

Abstract
Merely translating web sites is not sufficient for serving international clienteles. Web sites need to be “localised”. This involves adapting various informational aspects to address the local population in such a way that users understand the content and its use in the context of their own culture. A cultural marker denotes a convention used on a web site to address a particular population. Research in the area of localisation has concentrated on commercial web sites and software. We found that localisation of cultural web sites increases the complexity of the information management issues. As a project of the Section on Audiovisual and Multimedia of IFLA, a kit for localising the The Moving Image Collections (MIC) site was developed, then tested by using it to localise a selection of pages from the web site in French, Spanish, and Arabic. The kit, in the form of a .pdf file, can be used to produce a version of the MIC site localised for any other language or ethnic community.

Localisation of web sites
Merely translating web sites is not sufficient for serving international clienteles. Web sites need to be “localised”. This involves adapting various informational aspects to address the local population in such a way that users understand the content and its use in the context of their own culture. A number of definitions of localisation (also called localization, l10n, internationalization, i18n, globalization, g11n) are available. One definition that is easy to understand is the following: “the process of adapting technical media products... into a form where they are linguistically, functionally and culturally acceptable in countries outside the original target market. These media products take on a variety of different forms, e.g., software, help, printed and on-line documentation, web sites, e-mails, web based applications, multimedia components (audio, video, graphics) and so on” (Maroto and de Bortoli 2001, 4). The way the information is presented to users is important. In the case of localising web sites, cultural “markers” need to be identified and adjusted to each local situation. A cultural marker, “such as a national symbol, color, or spatial organization, denotes a conventionalized use of the feature in the web-site, not an anomalous feature that occurs infrequently” (Barber and Badre 1998).

A concept associated with translation and localisation is the notion of transcreation. The term refers to the idea that texts should not be translated word for word, but instead should be formulated in the target language in such a way that they read as if they had been written in that language in the first place. Transcreation is a kind of interpretation of a text and reformulating of it in another language (MultiCulture 2006, English and Spanish glossaries). Texts that have been transcreated have the important advantage of being much more intelligible to users than translated text. On the other hand, whether automatic transcreation can take place remains to be seen. After many years of research and development, the thorny problem of automatic translation has still not been resolved. However, more recent approaches to automatic translation look at language more holistically, manipulating larger chunks of text as units and thus reducing the possibility of putting words together in wrong combinations in the target language. One can imagine that such principles could be extended eventually to encompass transcreation as well. Meanwhile, the argument is often made that a text translated by software may be imperfect, but it is better than nothing. In fact, it may not be. Web users are famously intolerant of poorly designed sites, leaving them within a minute or two of their arrival at the site if it is too hard to use.
Depending on a particular user’s information needs at a particular time, a poor translation could well be the reason a user leaves a web site for another.

Critical to the issue of localisation is the question of varying points of view on a concept, person, phenomenon. A series of clever ads from the Hongkong and Shanghai Banking Corporation (HSBC 2008) drives home this point eloquently. In designing web sites (or anything else), other points of view than one’s own need to be considered.

Figure 1. Ads from HSBC that can be seen in airports and elsewhere.

Most of the research on localisation has concentrated on the question of conducting commercial activities via web sites. For example, Jagne (2004) studied cultural and social factors in shopping behaviour, and Boor and Russo (1993) studied associations of colours in various cultures. Other work is in the area of software localisation, for example Deitsch & Czarnecki 2001, Esselink 2000, Schmidt 2000. The question of localising cultural web sites has only begun to be explored. Key articles in this area include Barber and Badre (1998), Maroto and de Bortoli (2001), Elnahrawy (2001), Cunningham (2005), Duncker (2002), and Cleary (2000).

In our own work in this area, one study on this question (Bolduc, Genest and Leblanc 2004) was carried out by students at the Université de Montréal. A research project (MultiCulture 2006) was undertaken in 2005 to study multilingual retrieval of cultural content, and one part of the project involved studying the question of localising cultural content. Another research project was undertaken in 2006 to build on this groundwork to construct a kit for localising web sites with cultural content, and to test the kit on the web site of Moving Image Collections or MIC (2005). This project, called the MIC localisation project, is the focus of the present paper. In the course of doing the work, we built on the work done by the business community. Our results show that contextualising information for various cultural communities increases the complexity of the information management issues needed for knowledge organisation.
Moving Image Collections (MIC)

Moving Image Collections (MIC, pronounced “mike”) is an initiative of the Association of Moving Image Archivists (AMIA 2008), in collaboration with the Library of Congress. After the initial work was done in building the resource, the site was installed at the Library of Congress, where it now resides. MIC “documents moving image collections around the world through a catalog of titles and directory of repositories, providing a window to the world’s moving image collections for discovery, access and preservation” (MIC 2005). It seeks to build a union catalogue of moving image resources worldwide.

Institutions register as members, then load their records into the MIC database. In addition, MIC provides cataloguing tools and support for archives lacking resources, especially small archives, for example by offering an online form to help them catalogue their records using recognised standards and loading the records directly into the database. The site also includes an Archive Directory, which offers information about individual collections, as well as information on preservation of moving images and sound, programming for exhibition, and support for science educators who wish to use moving images in the classroom.

At present almost all the information on the MIC site is in English. While many users from other language communities may understand enough English to be able to use the site to some degree, versions of the site in other languages will be much more helpful to speakers of other languages, and ultimately will enlarge the user base as collections from around the world are included in the MIC database. We tried to identify what some of the needs of those users might be.

The MIC localisation project

After an initial study to try to get an understanding of the information needs of users from a number of different language communities, the Section on Audiovisual and Multimedia (AVMS) of the International Federation of Library Associations and Institutions (IFLA) adopted the MIC localisation project as a Section project at the Buenos Aires conference in 2004. Funding was approved after the Oslo conference in 2005. Work began in 2006, and was completed in 2007. The objectives of the MIC localisation project were to advise the MIC team on the needs of non-English-speaking users, build a kit for localising the MIC site, and test the kit by localising a selection of pages from the site in French, Spanish, and Arabic.

Planning meetings were held in Montréal in July 2006. These included demonstrations and discussions on localisation, a review of the entire MIC site, choosing the pages to test in the context of the project, and discussing the content of the localisation kit. In addition, plans were made for hiring and instructing student assistants, ensuring coordination of the work, and assuring quality control once the work was completed. At these meetings, a timeline for completing the work was also constructed. Twenty-one pages from the MIC site were chosen for localisation on the basis of their importance and usefulness to the international community.

The kit was constructed, the pages were localised in the three languages, and the files were delivered to the MIC Project Manager in the summer of 2007. Because of a number of constraints, including moving the Motion Picture, Broadcasting and Recorded Sound Division of the Library of Congress to a new facility in Culpeper, Virginia, the MIC team have not been able to install the files as of the spring of 2008. Once they are installed, it will be possible to evaluate their impact. Once this has taken place, the kit can be installed
on the MIC site. Anyone willing to do the work would then be able to download the kit (a pdf file), then construct localised pages of the MIC site in any language.

**Issues identified for localisation**

As we mentioned, the groundwork done on localisation has to do with using web sites for commerce. In addition to translation or transcreation of the text of a web site into other languages, factors that need adapting to other cultures include the design of the web site, its layout, navigation methods and technical characteristics. For cultural sites, interest in the literature is focussed on usability for multicultural communities (for example Barber and Badre 1998). Localisation is only considered implicitly.

Design of a single site to accommodate a number of different user communities is not the same problem as designing multiple versions of a site to accommodate various user communities. For a number of reasons, not the least of which is usability, multiple versions of a site is by far the better option because each version is addressed to a target group of users, who do not have to wade through information directed at other communities to get to the information directed at their own.

A number of markers such as formats and syntax for dates, times, addresses, postal codes, telephone numbers, units of measurement, representation of large numbers, and currencies are common to many web sites and are not specifically cultural issues. Other factors such as images, colours, and the order of elements of people’s names have cultural meaning attached to them. Symbolism or association with animals, national flags, situations, configurations, and so on, are also largely cultural in nature and require more attention and a more subtle analysis in making decisions about how to represent information. Metaphors existent in the local culture, various ways to represent a concept, and the choice of examples to use represent another level of subtlety that needs consideration. Here we explain some of these issues in more detail.

**DATE FORMATS.** A number of calendars are used around the planet, but even within countries that use the Gregorian calendar, a number of formats are used to express the date. For example, June 28, 2004:

- **United States** ....................... 6/28/2004
- **Most countries** .................... 28/6/2004
- **International system (SI)** .... 2004/06/28

Expressing the year with only two digits is the source of endless trouble, especially in the first decades of this new century, and especially in countries like Canada, where all three systems are in use. Traditionally, Canada has expressed the date like most countries, in the format of date-month-year. However, many computer systems, especially those used in commerce, are purchased from the U.S. without requiring an adjustment, so that often the date appears in U.S. format (month-date-year), for example on a bill or a receipt. More recently, there is a move to adopt the international system (SI, year-month-date). However, while all three systems remain in use, it is difficult to parse the date when the year is expressed in two digits. We will have to wait until 2032 (when the maximum number of days in a month is finally surpassed) for this ambiguity to be removed. At present, confusion reigns. For June 7, 2004, we get:

- **United States** ....................... 06/07/04
- **Most countries** .................... 07/06/04
- **International System (SI)** .... 04/06/07
A user accustomed to the American system would read the second date on the list as July 6, 2004 and the third as April 6, 2007. Whatever system is used, the year should always be expressed with four figures, in order to help users interpret the dates correctly. Indeed, the Long Now Foundation (1996) advocates the use of five. Thus localising the date format means displaying dates in the format used locally by a particular user group. Note that the punctuation also varies. The slash (/) can be replaced by a space, a non-breaking space, a hyphen (-) or a period (.)

**Representation of large numbers and money.** Several systems for representing numbers are in use. Groups of 3 digits representing thousands are variously punctuated with commas, periods, or nonbreaking spaces (3,242,672 or 3.242.672 or 3 242 672). In addition, there is confusion among languages about large numbers (million, milliard, billion, trillion). Even within English, these concepts are subject to interpretation. In the United States a billion means a thousand million (1 000 000 000), while in the United Kingdom it means a million million (1 000 000 000 000); similarly, in the U.S. a trillion means a thousand billion, and in the U.K. a billion billion. Representations of quantities of computer memory, derived from binary calculations, are rounded in many contexts, but may need to be exact in other contexts. Thus 1 GB or 1000 MB is more exactly 1024 MB.

For representing money, usage also varies between countries. The dollar sign ($) is used to represent many currencies called dollars (e.g. U.S., Canada, Australia) and in many other countries to represent pesos or other currencies. Sometimes the sign goes before the number, sometimes after. Even within Canada, an amount is expressed differently, depending on whether the context uses English ($35,244.87) or French (35 244,87$).

**Expression of names.** How the names of people are expressed is important because it relates to their identity. Depending on the culture and the context, names can appear as Firstname Lastname or, more often, the inverse. However, the question is far more complex than just this issue, because both first names and last names often have two or more elements. There are places where either the father’s or the mother’s last name is used, or both are used, and places where children’s names are derived from the father’s or mother’s first name with “son” or “daughter” added. When users are unfamiliar with various given names and family names, there is much room for confusion about which is which. When people immigrate to a country that uses another system, sometimes they adopt the practice of expressing their name in that country’s format, but sometimes they do not.

**Animals.** Animals, especially mammals, are loved universally, perhaps especially by children. However, that does not stop humans from killing them for food. A number of animals represent deep-seated associations in different cultures, sometimes revered for various desirable qualities they may have, sometimes reviled for their undesirable qualities in relation to a culture, and sometimes seen simply as lunch. Ready examples of animals associated with strong emotions are horses, rats, snakes, various kinds of birds, tigers, dogs and cats, and insects. Those responsible for localising web sites have every interest in studying the implications of including animals in representations to various populations.

**Colours.** These are associated with deep connotations among various cultures, although within a culture individual reactions abound. It is important in web site localisation to study the implications of the colours one chooses, in order to avoid gross errors in ad-
dressing particular cultures. Fortunately, information on the subject is readily available, although it is sometimes conflicting. However, in our work on this question, we found that apart from the most obvious problems (such as black for mourning in the west, white in the east), the design guidelines having to do with contrast and readability are much more important in choosing colours than cultural associations they may have.

**Choice of Images.** Iconic images are often used for illustrations on cultural web sites. Although some have universal appeal, others need to be localised to help the target user population make the desired association, otherwise the association is missed.

**Examples for Explaining Procedures.** Our kit uses the example of cataloguing a film, such as might appear on the MIC site. In presenting examples to help users understand various concepts, it is helpful to give local examples, since users are already familiar with them, and thus can focus on what the example is trying to illustrate. In the United States, *Citizen Kane* is a good film to use as an example in explaining how to create Dublin Core metadata for a film, because everybody knows this film. However, for an Italian audience, *La Strada* would be a better example, and for a Chinese cataloguer, *The dividing wall*.

Cultural markers go beyond the question of merely formatting information in a variety of ways to address various populations. They require much more abstract analysis of the information content, and involve issues of judgement, appropriateness, point of view, aesthetics, and symbolism. Communication takes place on many levels, and cultural markers are found on a conceptual level. Thus automating the transcreation and localisation of cultural web sites offers a challenge beyond that of automating commercial sites.

**Discussion and Conclusion**

The kit we developed for localising the MIC site takes into account the markers discussed here, and provides explicit instructions for localising the selected pages from the site to the research assistants. In addition, the test version of the kit contains an instruction to the assistants to make notes on using the kit, so that we can make any necessary adjustments before installing it on the MIC site for general use. By making the kit available, we make both a research contribution and a practical contribution to providing localised web sites for digital cultural content.

In time and with more work, eventually all cultural markers could conceivably be identified, along with their roles in specific cultures. Of course, these are subject to change, ironically largely because of globalisation. It is desirable to automate the localisation of web sites as much as possible, along the lines of generating web pages dynamically from databases of information. The most obvious differences in information elements from one language/culture combination to another can be addressed by good data structures and good programming. Data structures can atomise the data elements, and programming can then generate the correct localised version of some elements, such as formats for dates, time, money, and the syntax of addresses and telephone numbers. Cultural markers tend to be more abstract and involve human intervention and design decisions, although some ways of automating the expression of these might be developed as well. Although it may sometimes seem that we’ve always relied on the web for seeking information, it is still very young. The issue of localising web sites needs much more study and practice before we have a truly world-wide web, but we can hope to see such a beast in the foreseeable future.
Acknowledgements

SSHRC funded our work on localisation of cultural sites, and IFLA funded the work on the MIC localisation project through its FAIFE committee. Bruce Royan, former chair of the AVMS, led the initiative for obtaining the funding from IFLA. Anes Essid was a research assistant on the localisation aspect of the MultiCulture project. Samira Sambaíno, of the Instituto Universitario Centro Latinoamericano de Economía Humana (CLAEH) in Montevideo supervised the work on the Spanish-language localisation, and Marwa El Sahn, of the Biblioteca Alexandrina in Alexandria supervised the work on the Arabic localisation. To all these we are grateful.

References

A Time-aware Ontology for Legal Resources

Abstract
This paper presents a new approach to associating metadata to legal documents by exploiting a fully developed information ontology of legal resources that takes time into account. Our information ontology is technically an application of the FRBROO model to organization of legal documents. Our contribution clarifies not only the diachronic evolution of the legal resources in time, but it also puts the theoretical grounding for the modeling of the relationships between the different entities participating to the legislative process workflow (e.g. bills, amendments). Our model is also applicable to all artefacts of the publishing process. Moreover the time dimension can be used to support successful interconnections between different legal resources (e.g. between normative acts and case-law) that need precise point-in-time referencing.

Introduction
Document models based on the IFLA Functional Requirements for Bibliographic Records (IFLA, 1998) are used within several existing projects in computer support of activities in the legal and legislative domains (Lima, 2007) such as the Akoma Ntoso (Vitali, 2007), LexML Brasil, CEN Metalex (Vitali et al., 2007) and Norme in Rete (Archi et al., 2000) projects. Yet, they are mostly oriented to bibliographic organization of documents without a real modeling of the peculiar characteristics of the legal domain. Also, all of them refer to the current official version of the FRBR model, called FRBRER. Recently the FRBR model has undergone major revisions and a new version using an object-oriented approach is being developed.

Several legal ontologies exist concerning the legal resources but are mostly aimed at modeling the content, like LKIF-core (Breuker et al., 2007), or the issues connected to the legal language, like LOIS (Peters et al., 2007). Yet, we are currently missing an ontology to represent legal resources themselves, especially considering the information perspective and the time dimension. We consider this a preliminary step to a full modeling of legal knowledge.

From FRBRER to FRBROO
Initially described as entity-relationship model to the organization of bibliographical records, FRBR is being revised by three working groups. One of these groups is working on the harmonization of FRBR concepts with the CIDOC CRM ontology (ICOM 2004). This integration process between the main reference models of library and museum communities, started in 2003 and not yet finalized, has been a “good opportunity to correct some semantic inconsistencies or inaccuracies in the formulation of FRBR” (Bekiari, Doerr & Le Bœuf, 2008, p. 9). This harmonization effort also introduces the dimension of time in the FRBR model, which is essential to the museum community as well as to the legal domain.

According to Smiraglia (2003) the concept of work became central for retrieval leaving the role as secondary entity in the first catalogues, which focused on the inventory function (where item as central), having now a more important role in modern catalogues after noticing that users of a information retrieval system are interested in the content and
not in a specific manifestation. Even before the publication of the FRBR<sub>ER</sub> model, some research argued about the need of the creation of an entity which grouped works derived from another work. For example, Yee (1995) and Svenonius (2000) defended the creation of the entity “Superwork”. Other researchers, which defended the essence of the same idea, nominated this entity “Bibliographic Family” (Smiraglia & Leazer, 1999), “Textual Identity Network” (Leazer & Furner, 1999) and “Instantiation Network” (Smiraglia, 2002).

The new FRBR<sub>OO</sub> crystallized the results of researches and proposals regarding the Work, the most abstract concept of the FRBR<sub>ER</sub>, creating the new class “F15 Complex Work” which allows the grouping of Works according to some criteria. The “F14 Individual Work” class has been defined to model the associated concepts with a specific group of signs (F22 Self Contained Expression). The new model defines seven other classes related to the entity Work (F1 Work, F16 Container Work, F17 Aggregation Work, F18 Serial Work, F19 Publication Work, F20 Performance Work, F21 Recording Work). Apparently complex, this new more detailed modeling allows us to represent each aspect which in FRBR<sub>ER</sub> were under the umbrella of a single entity (the Work Entity) making the model easier to use.

FRBR<sub>OO</sub> and Legal resources
In order to apply the new FRBR<sub>OO</sub> model to legal resources, we give the following operational definitions of the main legal concepts:

- **Norm** — A rule of conduct issued by a competent authority and prescribing or regulating behaviour among individuals and within society. Its form of expression may be the written or the spoken word, but it may also be visual or be based on usage and custom.

- **Normative provision** — Any group of words or piece of writing expressing a norm or series of norms.

- **Normative document or act** — An officially published legislative written document through which a competent authority brings a norm into being.

- **Legal system** — A set of norms belonging by some criterion to a single system and related to one another in different ways, as by hierarchy (one norm having a higher or lower standing than another), generality (more specific or more general), time (issued before or after another norm), and modification (one norm modifying the other norm or getting modified by it).

- **Normative system** — The same legal system viewed from the outside is **dynamic**: it changes over time and can be represented in its evolution as a series of snapshots or film-stills in succession. The sequence in the time of legal systems so captured we will call the normative system. (Palmirani, 2005)

Figure 1 shows the “F27 Work Conception” and “F28 Expression Creation” events that create an original document (F4 Manifestation Singleton), as well as the “F32 Carrier Production” event that forms a “F19 Publication Work” which, in turn, produces various items. Normally the “Act 1 Text” and “Act 1 Published Text” instance has the same content (set of signs), but sometimes it is necessary to publish official communications with rectifications.

Table 1 relates classes and instances which can be identified in a signed official normative document.

A specific edition of an official gazette contains various entities which coexist. On a first analysis it is possible to perceive the publication per se which is the result of an industrial process and generates a number of issues according to a determined production plan. If we abstract the publication, it is possible to perceive the entities related to the shown legal resources. Table 2 relates classes and instances which can be identified on a page of an official publication.
During the life cycle of a legal norm various events can affect its content in relation to its form as well as to its subject matter. For example, the normative expression of a norm could be affected by actions of integration, modification or repealing. In the moment of an event of a norm modification, a derived work is created and it is represented by a new instance of “F14 Individual Work” class with the respective instance of “F22 Self Contained Expression” class. It is important to point out that the creation of a modified norm (derived) occurs in the “in force” date of the modifying norm. These dynamics are illustrated by Figure 2 which shows what happens when a norm (“Act 1") is modified by another norm (“Act 8”).

Table 1. Original Signed Act (Author Contribution)

<table>
<thead>
<tr>
<th>Class</th>
<th>Instance</th>
</tr>
</thead>
<tbody>
<tr>
<td>F4 Manifestation Singleton</td>
<td>The original signed document is a physical object that carries an instance of F22 Self Contained Expression.</td>
</tr>
<tr>
<td>F14 Individual Work</td>
<td>The concepts associated with the signed act document.</td>
</tr>
<tr>
<td>F22 Self Contained Expression</td>
<td>The normative provisions resulting from legislative process.</td>
</tr>
</tbody>
</table>

Table 2. Official Gazette Issue (Editor/Publisher Contribution)

<table>
<thead>
<tr>
<th>Class</th>
<th>Instance</th>
</tr>
</thead>
<tbody>
<tr>
<td>F5 Item</td>
<td>The Official Gazette issue (physical object)</td>
</tr>
<tr>
<td>F3 Manifestation Product Type</td>
<td>The publication product type “Diário Oficial da União” issued in 6th August 2007 by Brazilian Official Press.</td>
</tr>
<tr>
<td>F18 Serial Work</td>
<td>The periodical entitled “Diário Oficial da União”</td>
</tr>
<tr>
<td>F19 Publication Work</td>
<td>The concepts associated with the official publication issued in 6th August 2007 by Brazilian Official Press.</td>
</tr>
<tr>
<td>F17 Aggregation Work</td>
<td>6th August 2007 by Brazilian Official Press.</td>
</tr>
<tr>
<td>F41 Publication Expression</td>
<td>Complete layout and content provided by a publisher (including table of contents and expression from other works).</td>
</tr>
<tr>
<td>F22 Self Contained Expression</td>
<td>The normative provisions published.</td>
</tr>
<tr>
<td>F14 Individual Work</td>
<td>The concepts associated with the published act.</td>
</tr>
</tbody>
</table>
The concepts of Normative and Legal Systems are implemented in our model using instances of class “F15 Complex Work”. The set of legal resources which exist on a determinate date make a Legal System. The set of Legal Systems make a Normative System. This systematization is illustrated in Figure 3 which shows the elements of the previous example (Figure 3) grouped into Legal and Normative Systems.
The publication of a normative text can be accompanied by complementary information in sections normally named “Annex”. This kind of relationship can happen in a recursive way, that is, an annex can have other annexes. Using the Ontology of Universals (Guarino 1999) terminology, an annex is considered a Role and not a Type. As Roles have a limited organizational relevance, to represent an annex it is necessary to analyze the annex content. In the case that this entity has its own identity criteria such as regulations or international treaties, the annex is classified as “independent annex”. In the other case, when the entity presents complementary information and which are dependent of the main part, such as tables, we classified it as “dependent annex”.

When an independent annex exists, an instance of class “F15 Complex Work” should be used as a way of grouping instances of “F14 Individual Work” which exist to each component (main part and annexes). In the case of dependent annexes, the property “R5 has component (is component of)” is used between “F22 Self Contained Expression” instances and there is no need, in this case, to define additional instances at Work level. Figure 4 illustrates the two situations described above.

Figure 4. Norm and Annexes

When a norm is published in different languages, in cases such as legal translations or simultaneous publications in multiple languages, even though each text has the same value on a legal perspective, the legal resources are modeled with distinct instances of class “F14 Individual Work” and the respective instances of “F22 Self Contained Expression”.

As class “F22 Self Contained Expression” does not have the property which associates its group of signs to a specific language it is necessary to associate the entity represented in “F22 Self Contained Expression” instance with one “E33 Linguistic Object” instance and respective “E56 Language” instance. This is possible because FRBRoo is considered, using the terminology of Masolo et al. (2003), a multiplicative ontology, allowing colocalized entities.

Figure 5 represents the modeling to norms published in different languages with the same legal validity.
Conclusion
The proposed model describes the evolution of legal resources in time and could model the relationships between the entities participating to the legislative process workflow (e.g. bills, amendments). Moreover the time dimension successfully expresses the interconnections between different legal resources that need precise point-in-time referencing (e.g. between normative acts and case-law). Our model correctly describes all actual artefacts of the publishing process, such as legislative gazettes and other publications, including re-publications and errata.

This paper shows a model of the information objects of the legal domain and in particular focuses on the evolution of legal resources in time, defines the relationship of normative and legal systems concepts, and provides explicit support for modeling annexes and norms published in more than one official language.

Acknowledgement
This research was partially supported by ESTRELLA project CEC-FP6-IST Programme Contract Number: 027655 and by Brazilian Government under CAPES Grant BEX 4414/06-3.

References


Design and Development of a Bilingual Thesaurus for Classical Tamil Studies
Experiences and Issues

Abstract
Based on experiences with the design and development of a Tamil-English bilingual thesaurus this paper discusses with examples: (1) some issues related to vocabulary management in multilingual thesauri in culture-specific domains; (2) special aspects of the Tamil language in this regard; (3) alternative ways of linking certain descriptors to lengthy lists of NTs and RTs; (4) advantages of integrated use of two or more knowledge organization tools; and (5) use of the bilingual thesaurus for certain types of research in Tamil. Issues related to equivalence, non-hierarchic associative relationships, homographs, NTs are discussed. The paper suggests that integrated use of two or more knowledge organization tools adds value.

Introduction
Bilingual and multilingual thesauri are useful tools in cross language information retrieval that require translation of queries and document representations from input language to other languages. Jorna and Davies (2001) have remarked that: “… multilingual tools are gaining importance as increasingly diverse groups from different cultural and linguistic backgrounds seek access to equally diverse pieces of information”. Even within the Web community, there is growing interest in vocabulary-based techniques. Harpring (1999) gives an overview of the Getty’s vocabularies with examples of their use in Web search interfaces.

Recently, the Government of India established a Centre of Excellence for Classical Tamil (CECT) to promote ‘the cause of Classical Tamil’. The Centre has initiated a project to build a digital library of Tamil classics of the Sangam period1 (Sharada and Manju Naika, 2007) and, as part of this, a Tamil-English bilingual thesaurus covering the domain of ‘Tamil Studies’. Some issues relevant to multilingual thesauri in culture-specific domains have been reported in another paper. (Neelameghan and Raghavan, 2005) This paper discusses the following aspects of multilingual thesaurus construction based on the experience in working on this ongoing project:

• Vocabulary management in bilingual and multilingual thesauri in culture-specific domains;
• Special aspects of the Tamil language in this regard;
• Alternative ways of linking descriptors to lengthy lists of NTs and RTs;
• Advantages of integrated use of two or more knowledge organization tools; and
• Use of the bilingual thesaurus for certain types of research in Tamil.

The Approach
Tamil, a Dravidian language with its own script, is independent of both Sanskrit and Latin and has a literary tradition going back to around 300 B.C. The domain of Tamil Studies covers a wide range of subjects including practically all areas of humanities, many social

1. Many historians refer to the Tamil literature from ca. 300 B.C. to 300 A.D. as Sangam literature; Sangams were Tamil academies, which enabled poets and authors to gather periodically to publish their work (Wikipedia).
sciences and a few sciences. As there is no information retrieval thesaurus covering the domain, the thesaurus is being built using a ‘bottom up’ approach starting with a corpus of 125,000 Tamil terms identified using available lexical tools. The principal source has been the Tamil Lexicon of the University of Madras (1924–1936), web editions of which are available (University of Chicago, Dictionaries of South-Asian Languages (DSAL); University of Cologne, Tamil Lexicon (COTL)). In building the thesaurus a modular approach based on the schedule of main classes of Colon classification was employed. This helped to have adequate control, even though integrating records for terms in different disciplines could raise problems of handling duplicate entries at a later stage. The features of the thesaurus are described in another paper (Neelameghan and Raghavan, 2007). The thesaurus is being maintained as a WINISIS database with the following fields:

<table>
<thead>
<tr>
<th>Tag</th>
<th>Name</th>
<th>Repeatability</th>
<th>Sub-fields</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Descriptor</td>
<td>abcde</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Descriptor1</td>
<td>a</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>SN</td>
<td>R</td>
<td>a</td>
</tr>
<tr>
<td>3</td>
<td>US</td>
<td>a</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>UF</td>
<td>R</td>
<td>a</td>
</tr>
<tr>
<td>5</td>
<td>BT</td>
<td>R</td>
<td>a</td>
</tr>
<tr>
<td>6</td>
<td>NT</td>
<td>R</td>
<td>a</td>
</tr>
<tr>
<td>7</td>
<td>RT</td>
<td>R</td>
<td>a</td>
</tr>
<tr>
<td>900</td>
<td>Links</td>
<td>abcdedef</td>
<td></td>
</tr>
</tbody>
</table>

The Field 11, Descriptor1 is being used for inputting the descriptor in a format required for hyper-linking with the DSAL database. Field 1 and all the words (excluding apparatus words) in field 2 are indexed. Terms in all the fields except fields 2 and 4 are hyper-linked either to the corresponding thesaurus record or to DSAL’s Tamil Lexicon. Some of the issues relevant to the bilingual thesaurus are discussed below.

**Script and Transliteration**

Tamil terms were transliterated and entered in Roman script. The major concern was to adopt a transliteration standard that supported automatic and consistent conversion to Tamil script. The COTL transliteration scheme was adopted as it met these requirements. In other words terms in the Tamil script could be added as a separate field to the database at a later date without having to key-in the terms.

**Equivalence Relationship**

A language is a product of, and reflects the culture of a community. It is the culture and lifestyle of the members of a community that necessitate the formation of lexemes/expressions (words/terms). Concepts in culture-specific domains have meaning in the life of the members of the community belonging to the culture. Unless communities speaking two different languages share the same culture, it is likely that certain concepts in culture-specific domains may have unique verbal expressions only in a particular language. For example:

- cAttunARRu = Young plants planted in place of the dead ones
- aSTAgkaputti = Eight Kinds of Knowledge
- cARvAkam = cAruvAka’s materialistic philosophy which says perception is the only source of knowledge
In this thesaurus two issues relating to equivalence relations had to be handled:

- The presence of a large number of synonyms in Tamil for many of the concepts
- The difficulty in finding equivalent concepts (and terms) in the English language

The notion of ‘equivalence relationship’ had to be extended to include both synonyms in Tamil and terms in English denoting near equivalent/semantically close concepts. In practice different strategies, depending on the situation, were employed:

- To use the nearest semantically equivalent English term;
- To use an appropriate phrase/compound term in English that represented the concept;
- To use the original term itself in case of concepts almost accepted as proper nouns.

The second issue related to ‘equivalence relationship’ was the occurrence of several terms in Tamil with closely similar connotations. Searches for two synonymous Tamil words, ‘tAmarai’ and ‘kamalam’ and their English equivalent, ‘lotus’ carried out in the online Tamil Lexicon (http://dsal.uchicago.edu/dictionaries/tamil-lex/) retrieved records in which the input string occurred either as an independent word or embedded in another word. More than 200 of the 327 terms were related to ‘tAmarai’ and 82 of these were synonyms!

**Table 2. Multiplicity of Synonyms**

<table>
<thead>
<tr>
<th>Search Term</th>
<th>No. of Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>tAmarai</td>
<td>327 Entries with tAmarai as entry word or in the explanation</td>
</tr>
<tr>
<td>kamalam</td>
<td>36 entries with kamalam as entry word or in the explanation</td>
</tr>
<tr>
<td>Lotus</td>
<td>309 entries with Lotus as entry word or in the explanation</td>
</tr>
</tbody>
</table>

**Homographs**

Many of the Tamil terms were found to be homographs. For example, ‘tAmarai’ has been used to indicate:

- Lotus plant;
- Lotus flower;
- Lotus as a shape (entities in the shape of a lotus);
- Lotus-like properties (e.g., soft like lotus petals)

Some other examples of homographs in Tamil are:

- *appu* = Thigh; Father; Loan; Debt; Domestic male servant; Water; Trumpet tree; Sixth division of day
- *iTimpam* = Baby; Misery; Spleen; Egg of birds; Castor plant.

The evolution in the meaning and connotation of terms in Tamil probably explains the presence of a number of homographs. The Tamil words *kurinchi, mullai, marutam, neitl,* and *palai,* for example, denoted the plants, creepers, trees (all flora) found in abundance in the region concerned. Over a period of time the regions where these plants were found to grow came to be referred to by these same terms. Subsequently even the lifestyle characterizing the region came to be referred to by the term. This probably explains the use of the names of the regions for the codes of life prevalent in various regions. (Neelameghan, 2008). The real meaning is to be understood in the context. In building the thesaurus it became necessary to extensively use role operators to indicate the different connotations and contexts of the Descriptor. For example:

- *iTimpam* (baby); *iTimpam* (castor); *iTimpam* (egg); *iTimpam* (misery); *iTimpam* (spleen)
Hierarchical Relationship

The general guideline adopted for hierarchical relationships between concepts was to treat types or varieties as also parts of an entity as narrower terms. In reality a part may be part of any level (or order). The example below represents hierarchical relationship:

Tāmarai (lotus)
MirunALam (Stalk of the Lotus)
TāmaraimuL (thorny portion of the stalk of the lotus)

Parts of lower levels are labelled as NT1, NT2, NT3, etc. The major issue in handling hierarchical relationship arose from the occurrence of a large number of NTs to many descriptors; e.g., utti (literary device) has over 32 NTs; cangka-nU (Tamil classics) has 41 NTs, AmaNakku (castor plant) has 55 NTs, nAyAr (Saiva saints) has 63 NTs, ceyyuLLami (figure of speech) has over 180 NTs. The usual practice is to list all NTs, and to treat NTs to the same descriptor as RTs to each other. For example:

AmaNakku (castor plant)
SN Castor plant; Ricinus communis
BT tAparavastu (plant)
NT acAram (castor plant)
NT amanTalam (castor plant)
NT amanTam (castor plant)
..... ..... ......

Clicking on a NT, say acAram (castor plant) will link to the record:

acAram (castor plant)
SN Castor plant variety
BT AmaNakku (castor plant)
BT2 tApara-vastu (plant)
RT amanTalam (castor plant)
RT amanTam (castor plant)
..... ...... ......

This approach resulted in long schedules and a large number of linked entries. To reduce the size of the thesaurus without compromising on the links (e.g. to the related varieties of castor plants, in the above example), in all cases involving five or more NTs, the full list of NTs is given in one schedule under the descriptor with hyper-links to this record from each of the NTs. For example, the record for acAram (castor plant) as descriptor will be:

acAram (castor plant)
SN Castor plant variety
BT AmaNakku (castor plant)
BT2 tApara-vastu (plant)
For other varieties of castor plant see AmaNakku (castor plant)

Lateral Relationships (Non-Hierarchical Associative Relationships)

Conceptually lateral relationships (LRs) have remained a grey area. Theoretically a concept may have conceptual relationship with any other concept. However, from a pragmatic point of view it is both necessary and helpful to identify conceptual relationships that have the potential to contribute to enhancing information processes. While it may be difficult to measure the impact of providing for lateral relationships on retrieval effectiveness, it is generally conceded that associative relationships are important in a thesaurus to support navigation and for enhancing search. In reality many terms are associated with
any given term. Which of the LRs should be handled by the vocabulary has been the subject of discussion. A general rule could be that all near-permanent relationships between concepts should be handled by the vocabulary. A few efforts at defining and categorizing LRs have been reported (Neelameghan & Ravichandra Rao, 1975; Neelameghan & Maitra, 1978; Neelameghan, 2001). However, the schema was applied in the design of a substantive thesaurus only recently (Moholt, 2001). A comprehensive attempt to examine the semantics of relationships in knowledge organization has also been reported. (Bean, 1996; Bean & Green, 2001) The work related to the present bilingual thesaurus was used to identify LRs in culture-specific domains. A taxonomy of LRs between concepts has been developed. (Neelameghan & Raghavan, 2005; Neelameghan & Raghavan, 2006) An earlier version of this taxonomy has been mapped to their nearest equivalents in FrameNet highlighting the characteristics and also the varying degrees of compatibility between the two schemes. (Green & Bean, 2006) Working on the present thesaurus has also helped in continuously evaluating the schema. For example, one kind of LRs that had to be handled was:

- Transformation/Metamorphosis: Relationship arising from a change either in the external form or internal nature of an entity: Transformation of an individual following a mystical experience; Transformation of an individual after adapting another way of life (e.g., grihastAsrama to sanyAsAsrama); Telekinesis; Resurrection, etc.

Theoretically, a case for explicitly marking different kinds of LRs in a thesaurus could be made. However, at this stage, it is not clear as to how these could be used for enhancing navigation and retrieval; small scale experiments are on to obtain a feedback from the users on the utility of specifying the nature of LRs between a pair of concepts (terms) instead of merely labelling all as RTs. Such an approach could lead to:

- Improved precision in IR;
- Better mapping of a semantic domain; and
- Knowledge discovery.

Web of Relationships and Value Addition

The building of the thesaurus led to the recognition of a complex network of relationships between concepts. Given this, the near impossibility of providing, for any given concept, the complete network of relationship as also a full description was realized. For handling the web of relationships two other options have been explored:

- Maintaining especially long lists of NTs and RTs as separate files and hyper-linking the corresponding thesaurus record(s) to these files;
- Linking thesaurus records to appropriate lexical tools on the Web. Every record in the Tamil-English thesaurus has been linked to the page containing the corresponding lexical record in the online Tamil Lexicon (DSAL) (Figure 2).

Discussions with subject specialists identified some other special requirements; e.g. the need for listing, as descriptors, titles of all the 41 Tamil classics of the period, commentaries on these, etc. In handling this it was realized that listing of a large number of titles as NTs in an alphabetical sequence is not helpful. Instead grouping the NTs in a logical sequence as in the schedule of classics in Colon Classification was found to be more helpful to users.

Neelameghan & Parthasarathy (2008) based on their study of literary devices (utti) in Tamil scholarly communication show that knowledge organization tools (KOTs), such as thesauri linked to lexicon, can also aid knowledge discovery and researches in the subject domain covered by the KOT.
Conclusions

There is increasing awareness of the dynamic and context-dependent character of languages especially in the Humanities. This suggests the importance of tools providing conceptual information that explain the use and meaning of terms and also indicate their relationships to other concepts. Relationships are important for disambiguation in terms of showing the different meanings that terms can assume in a particular knowledge domain. Such tools are also of value in supporting formulation of well-structured queries that are known to yield better results than poorly structured queries. Bilingual and Multilingual thesauri especially in culture-specific domains raise several issues that need to be adequately addressed.
• It is helpful to use appropriate mechanisms and web-based lexical tools for adequate and comprehensive handling of relationships between concepts especially in culture-specific domains.
• It is difficult to find equivalent terms with one-to-one correspondence in a bilingual thesaurus especially in the Humanities.
• It is indeed difficult to implement parallel and identical hierarchies of related terms (BTs, NTs and RTs) for many concepts in two or more languages.

References
Indexing and Retrieving Images in a Multilingual World

Abstract
This paper presents the problem statement, the methodology and the preliminary results of a research project aiming to compare two different approaches for indexing images, namely: traditional image indexing with the use of controlled vocabularies, and free image indexing using uncontrolled vocabulary. The experiment intends to measure their respective performance for image retrieval in a multilingual context, in terms of effectiveness, efficiency, and satisfaction of the user.

Introduction
Images have always been a powerful communication tool. They present a multifaceted and actual value. They serve not only as a source of information but, with the development of more accurate visualization techniques, they also enhance the understanding of that information. For a few years, the diffusion of images has increased, mainly because of the development of digital technologies and the unprecedented growth of the Web. Confronted to this profusion of images, individuals now speculate on how to retrieve images with effectiveness and efficiency.

In general, two categories of queries are used to retrieve images on the Web: graphic or textual queries. In the first category, the individual submits a graphic query using an image or a drawing and the system tries to retrieve a similar image by using certain physical characteristics of the image such as color, shape or texture. However, the majority of image searches on the Web still use textual queries and the retrieval’s success depends on the correspondence between the query terms and the text (ancillary text or indexing terms) associated with the images. Since images do not always include a caption or any kind of text, the indexing process remains crucial. Image indexing has, so far, been divided between two camps: the proponents of “controlled vocabulary”, and the proponents of “uncontrolled vocabulary”. The former method focuses on indexing terms extracted from thesauri, classification schemes or subject heading lists, while the latter focuses on terms drawn from the natural language and does not impose any structure or mandatory element. This long-neglected form of indexing is currently becoming very trendy with the emergence of picture sharing web sites.

Image retrieval presents two major obstacles. Firstly, the process of transposing the content of an image into a verbal expression poses significant challenges to the individual. The second barrier comes from the “language” of the image itself. By their very nature, images are considered to be language-independent resources. Nevertheless, the text associated with the images gives the image a linguistic status similar to any other textual document, which can significantly affect its retrieval. And given the great linguistic diversity existing on the Web, we must expect that the text associated with the images exists in many different languages. For example, if a user formulates a query in English and the images to be retrieved are associated with English text, the cross-lingual problem does not arise. However, if the query language and the associated text of the image are different, the retrieval will not be possible, unless the retrieval system includes a Cross Language Information Retrieval (CLIR) mechanism which allows cross-language mapping between the query terms and the associated text.
Methodology
This research compared image retrieval within two contexts: a monolingual context, i.e., where the language of the query (French) is the same as the indexing language (French); and a multilingual context, i.e., where the language of the query (French) is different from the indexing language (English). For this research, three data collection methods are used.

Firstly, an analysis of the vocabularies used for image indexing was employed in order to examine the multiplicity of term types applied to images (generic description, identification, and interpretation) and the degree of indexing difficulty due to the image’s subject and nature. The objective of this analysis was to identify the specific characteristics of each indexing approach studied in this project. This analysis has been conducted on the indexing terms assigned to the 3,950 ordinary images drawn from a commercial online catalogue and included in the image database prepared for this research. Each image has been indexed in four different ways: with controlled vocabulary (French and English), and with uncontrolled vocabulary (French and English). A grid analysis was developed and applied on all indexing terms assigned to all images of the database. Three levels of analysis were determined: terminological, perceptual and interpretative. The terminological level refers to the lexicographical aspects of the indexing terms (number of indexing terms assigned, indexing term types, and indexing level). The perceptual level examines the attributes related to the physical, functional and identifying features of the described images. Finally, the interpretative level identifies more precisely the relationship between the words, with four categories of relations (generic, partitive, instance, and associative).

Secondly, a simulation of the retrieval process involving a subset of images was performed with 60 native French undergraduate students. For the retrieval simulation, 30 images have been randomly selected from the image database and successively shown to the participant who tried to retrieve them using a textual query in French. The quantification of the retrieval performance of each indexing approach was based on the usability measures recommended by the standard ISO 9241-11, i.e. effectiveness, efficiency, and satisfaction of the user (AFNOR, 1998, 2). For this study, effectiveness was defined by the number of retrieved images divided by the total number of images to be retrieved. For efficiency, a distinction has been made between temporal efficiency (measured by the time, in seconds, used to retrieve an image) and human efficiency (measured by the number of queries required to retrieve an image (Brangier and Bracenilla, 2003, 50).

Finally, a questionnaire was developed to collect the impressions of the participants concerning the retrieval process and the results obtained, in order to gather information on their satisfaction during and after the retrieval process and to complement the data collected during the retrieval simulation. Initially, the participants were asked to evaluate their degree of satisfaction regarding the retrieval results they got after each image searched. Secondly, a comprehensive questionnaire was presented to the participants, at the end of all the retrieval tasks. This questionnaire consisted of questions about their general perception of the retrieval tasks, their regular search practices and familiarity with image retrieval, and their demographic profile.

1. According to the Merriam-Webster Dictionary multilingual refers to “several languages”, several meaning “more than one” languages.
Results
This research had two main objectives. First, the indexing terms were analysed in order to characterize each of the approaches used to describe the ordinary images of everyday life objects contained in the database. Second, the simulation of a retrieval experience was conducted in order to evaluate which indexing approach is the most effective, efficient and satisfying for the image searcher. The partial results of the study are presented in the following paragraphs.

The analysis of all the indexing terms assigned to the 3,950 images of the database was carried out while involving a variety of attribute categories (terminological, perceptual and interpretative). The results of this analysis of the indexing terms of the four indexing approaches are summarized in Table 1.

<table>
<thead>
<tr>
<th>Level</th>
<th>Attributes Categories</th>
<th>Sub-categories</th>
<th>FCV %</th>
<th>FUV %</th>
<th>ECV %</th>
<th>EUV %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terminological</td>
<td>Number of Indexing Terms Assigned</td>
<td>1 Term</td>
<td>89.3</td>
<td>97.3</td>
<td>90.8</td>
<td>99.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 Terms</td>
<td>8.9</td>
<td>2.6</td>
<td>7.6</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 Terms</td>
<td>1.2</td>
<td>0.1</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 Terms</td>
<td>0.3</td>
<td>0.0</td>
<td>0.4</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 Terms</td>
<td>0.2</td>
<td>0.0</td>
<td>0.2</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6 Terms</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7 Terms</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Indexing Term Types</td>
<td></td>
<td>Single Terms</td>
<td>48.5</td>
<td>16.8</td>
<td>44.4</td>
<td>11.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Multiple Terms</td>
<td>51.5</td>
<td>83.2</td>
<td>55.6</td>
<td>88.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Complex Terms</td>
<td>0.9</td>
<td>8.9</td>
<td>0.6</td>
<td>4.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Abbreviations</td>
<td>0.9</td>
<td>8.2</td>
<td>1.2</td>
<td>10.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Neologisms</td>
<td>0.3</td>
<td>3.6</td>
<td>1.6</td>
<td>4.0</td>
</tr>
<tr>
<td>Indexing Level</td>
<td>Generic Terms</td>
<td>52.2</td>
<td>21.3</td>
<td>46.6</td>
<td>13.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Specific Terms</td>
<td>47.8</td>
<td>78.7</td>
<td>53.1</td>
<td>86.7</td>
<td></td>
</tr>
<tr>
<td>Perceptual</td>
<td>Physical</td>
<td>Form</td>
<td>1.6</td>
<td>3.3</td>
<td>2.2</td>
<td>6.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dimension</td>
<td>1.5</td>
<td>1.4</td>
<td>0.5</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Color</td>
<td>0.1</td>
<td>8.6</td>
<td>0.1</td>
<td>3.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Texture</td>
<td>0.4</td>
<td>0.2</td>
<td>0.2</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Material</td>
<td>4.7</td>
<td>16.3</td>
<td>3.0</td>
<td>9.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Body Part</td>
<td>2.8</td>
<td>1.1</td>
<td>2.3</td>
<td>2.1</td>
</tr>
<tr>
<td>Functional</td>
<td>Function</td>
<td>15.6</td>
<td>20.6</td>
<td>14.9</td>
<td>21.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Result/Product</td>
<td>1.7</td>
<td>1.0</td>
<td>1.5</td>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Activity/Action</td>
<td>7.6</td>
<td>13.1</td>
<td>7.6</td>
<td>16.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Localization</td>
<td>4.0</td>
<td>7.6</td>
<td>6.6</td>
<td>11.9</td>
<td></td>
</tr>
<tr>
<td>Identification</td>
<td>Trademark</td>
<td>0.0</td>
<td>25.1</td>
<td>0.0</td>
<td>6.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Product Name</td>
<td>0.0</td>
<td>1.6</td>
<td>0.0</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sponsor</td>
<td>0.0</td>
<td>1.6</td>
<td>0.0</td>
<td>1.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Geographical Place</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Other Attribute</td>
<td>Temporal</td>
<td>0.2</td>
<td>0.9</td>
<td>0.1</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Numeric</td>
<td>0.2</td>
<td>6.1</td>
<td>0.7</td>
<td>1.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Qualifier</td>
<td>11.5</td>
<td>7.7</td>
<td>10.7</td>
<td>6.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Person</td>
<td>3.8</td>
<td>2.3</td>
<td>3.1</td>
<td>7.4</td>
<td></td>
</tr>
<tr>
<td>Interpretative</td>
<td>Relations</td>
<td>Generic</td>
<td>46.7</td>
<td>60.7</td>
<td>54.8</td>
<td>82.9</td>
</tr>
<tr>
<td></td>
<td>Partitive</td>
<td>0.6</td>
<td>12.8</td>
<td>0.5</td>
<td>4.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Instance</td>
<td>0.2</td>
<td>28.3</td>
<td>0.3</td>
<td>8.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Associative</td>
<td>1.2</td>
<td>8.6</td>
<td>7.4</td>
<td>3.2</td>
<td></td>
</tr>
</tbody>
</table>
The analysis of the indexing terms revealed that the indexing form using French controlled vocabulary (FCV) distinguished itself from the other indexing forms on three points. First, on the terminological level, FCV assigned more indexing terms per image, and these terms tended to be single terms of general nature. Secondly, in terms of perceptual attributes, FCV did not offer a wide variety in the selection of attributes assigned to images. Finally, very few structural relations were encountered with this particular indexing form, in general.

The second indexing form examined uses French uncontrolled vocabulary (FUV). The terminological analysis of FUV showed that this form mainly differed from the other indexing forms with the attribution of complex terms (many different concepts contained in each indexing term). At the perceptual level, FUV is characterized by the use of many identifying attributes such as trademark, product name, sponsor and geographical location, and also physical attributes such as color and material. Finally, in terms of interpretative characteristics, FUV tended to offer more structural relationships compared to the other indexing forms.

The third indexing form we analysed is the indexing using English controlled vocabulary (ECV). This specific indexing form did not seem to distinguish itself by some attributes, unlike the other indexing forms. At the terminological level, ECV offered rather similar proportions of single/multiple terms and generic/specific terms. For the perceptual and interpretive aspects, ECV did not prevail on the other indexing forms with the use of specific attributes or structural relations.

Finally, the last indexing form studied here uses English uncontrolled vocabulary (EUV). At the terminological level, EUV mainly diverged with the lowest number of indexing terms used to describe the images of the database, and the highest number of multiple terms, specific terms, neologisms and abbreviations. At the perceptual level, EUV distinguished itself with the attribution of functional attributes such as function and activity/action, and physical attributes such as texture and dimension. Finally, EUV mainly outclassed the other indexing forms with a more important use of generic relations.

The second part of this research project involved a simulation of the retrieval process. A sample of 60 participants took part in the simulation. Each participant was randomly assigned to one indexing form, i.e. the queries used to retrieve the images were mapped to one of 6 indexing forms and the participants did not know to which indexing form they were associated. During the retrieval process, four main variables were registered by the retrieval system: the success rate of the retrieval task (image retrieved or not), the length in seconds of the retrieval task, the number of queries used to retrieve each image, and the evaluation of satisfaction once each retrieval task was completed. The results of the retrieval simulation are presented in Table 2.

<table>
<thead>
<tr>
<th>Indexing Form</th>
<th>Proportion of Retrieved Images (%)</th>
<th>Standard Deviation</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCV</td>
<td>69.0</td>
<td>11.66</td>
<td>10</td>
</tr>
<tr>
<td>FUV</td>
<td>76.0</td>
<td>18.78</td>
<td>10</td>
</tr>
<tr>
<td>FCUV</td>
<td>86.3</td>
<td>8.95</td>
<td>10</td>
</tr>
<tr>
<td>ECV</td>
<td>44.0</td>
<td>6.81</td>
<td>10</td>
</tr>
<tr>
<td>EUV</td>
<td>41.0</td>
<td>18.53</td>
<td>10</td>
</tr>
<tr>
<td>EUCV</td>
<td>51.7</td>
<td>8.46</td>
<td>10</td>
</tr>
</tbody>
</table>
The examination of the preceding results indicates that the indexing form using the combination of French controlled and uncontrolled vocabulary (FCUV) presents the highest average proportion of retrieved images (86.3%), while the indexing form using English uncontrolled vocabulary (EUV) obtained the lowest number of images retrieved (41.0%). Further analysis will be performed on the effectiveness of the indexing forms, including statistical analysis of variance with two factors, in order to determine whether the differences observed between the indexing approaches are significant are not. The two factors taken into account are the language used for indexing (two levels: French and English), and the form of indexation (three levels: controlled vocabulary, uncontrolled vocabulary, and the combination of controlled and uncontrolled vocabulary). Statistical analysis will also be conducted on the average time used to retrieve an image (temporal efficiency), and on the number of queries required to retrieve an image (human efficiency).

The averages of the satisfaction rates recorded by the participants are displayed in Figure 1. These rates are based on a Likert Scale, where “1” refers to the lowest satisfaction rate, and “5” to the highest satisfaction rate experienced by the participants, according to which indexing form their queries were mapped.

The examination of the preceding figure shows that the participants were more satisfied when their queries were mapped to an indexing form using French vocabulary. Moreover, it seems that the combination of controlled/uncontrolled vocabularies offered more satisfaction to participants, in French as well as in English. Statistical analysis will be conducted on these data as well.

**Discussion**

Overall, the preliminary findings of this study indicate that both studied approaches used for image indexing, with controlled vocabularies and uncontrolled vocabulary, present differences in their basic characteristics. The traditional approach using controlled vocabulary provides a consistent but less detailed description of images, while the indexing with uncontrolled vocabulary reveals an important diversity in the level of indexing and the variation of image attributes assigned to images. The comparison between both approaches confirms a fundamental difference which may have an impact on the image retrieval.

The use of controlled vocabularies for image indexing offers many advantages for retrieval, browsing and interoperability. The control offered by these vocabularies is mani-
fold. It manages the use of synonyms, homonyms, lexical anomalies, and so on. However, one of the main disadvantages of controlled vocabularies is that they quickly become outdated. For example, neologisms will often take a long time before they appear in controlled vocabularies. As a result, the search will be less accurate because controlled vocabulary will sometimes not allow a specific search. But the main difficulties associated with the use of controlled vocabularies for image indexing could be summarized as follows. First, the use of the majority of controlled vocabularies is beyond the capacity of the non expert or less trained professional. Secondly, and what is probably more crucial, these vocabularies are not suitable for all image types, and certainly not for the majority of ordinary images. Finally, most of the controlled vocabularies exist only in one language (in most cases, English) which implies that it will not be a great help in all linguistic contexts. Nevertheless, it seems that Web users created their personal solution to overcome the difficulties inherent to the use of controlled vocabularies for image indexing and use their own method to index images as they do with collaborative tagging, the latest trend in image indexing.

Collaborative tagging seems very attractive since it shows a close relationship with the real users and the way they see and describe things. Moreover, neologisms and all forms of newly-created terms are quickly integrated in collaborative indexing. Compared to indexing with controlled vocabularies and especially for new topics, collaborative indexing is likely to win hands down since the same words could take months, and even years, before they are even considered to be included in a controlled vocabulary. Besides, collaborative tagging can be done in one language or combine several languages, which means that the user does not have to know many languages to retrieve images indexed by users speaking different languages. However, despite its growing popularity and much like indexing with controlled vocabulary, collaborative indexing also presents several gaps. For example, some ambiguities emerge because the same keyword is often employed by several individuals, but in various contexts. In the same vein, the lack of synonymic control involves the use of many different keywords to describe the same concept. Consequently, free indexing employed in collaborative tagging systems is often considered to be of poor quality.

Conclusion and Future Work

The preliminary results presented above are the first step of an investigation on the use of different approaches for ordinary image indexing. Further analysis on the collected data is considered, including comparing query terms to indexing terms and examining translated terms to determine the influence of mistranslations. This is expected to help in identifying which indexing approach could better support retrieving images in a multilingual context.

The use of both indexing approaches is an old debate which was not fully resolved up until now (Arsenault, 2006, 142). According to circumstances, these two approaches may co-exist and be very helpful. Indexing with uncontrolled vocabulary does not have to be necessarily considered as an alternative or a replacement solution for traditional indexing with controlled vocabulary, but rather as a possibility for enhancement. The findings of this study suggest that in the near future, the information systems could benefit from allowing more and more the co-existence of controlled vocabularies and uncontrolled vocabularies resulting from collaborative image tagging for example, and giving the users the possibility to dynamically participate in the image indexing process, in a more user-centered way.
Acknowledgement
The author would like to thank the members of the Research Committee, Mr. Clément Arsenault and Ms. Christine Dufour. A special thanks to Mr. Philippe St-Germain who took the time to read this paper. This research is graciously supported by a grant from the Social Sciences and Humanities Research Council of Canada (SSHRC).

References
Knowledge Representation Focusing Amazonian Culture

Abstract
In this paper, cultural identity is shown to elaborate a language of information that has as its thematic universe the Amazonian Paraense culture. To accomplish this study of knowledge representation, cultural terms of popular histories were collected from residents of eight municipal districts of the State of Pará: Abaetetuba, Belém, Santarém, Castanhal, Bragança, Altamira, Cametá and Marajó’s Island. After defining the terms, this study elaborated the structure of the classification, establishing the relationships among the concepts, forming classes starting from your common characteristics. The initial structure of the classification, its categories, represents the knowledge in an organized way so that the information in a more agile way can be recovered. As result we created a classification of the thematic universe with the terms in agreement with the certain categories and, like this, to represent the knowledge of the rich and fascinating Amazonian culture, full of secrets and mysteries, in which the individuals produce a unique cultural environment.

1. Introduction
The Amazon is an area of Brazil rich on the fauna and flora, there Amazonian man lives in deep harmony relationship with that nature that surrounds him, where the rivers and the forests become very important for that daily one. Amid that atmosphere of great wealth and narratives, the project RESNAPAP (the Symbolic Representation of the Popular Narratives of Paraense Amazon as Language of Information), of Librarianship College of the Federal University of Pará (UFPA), in Brazil, began in 1998. It began as a subproject of the IFNOPAP project (The Imaginary in the Popular Orals Narratives Forms of Paraense Amazon), of the Letters and Communication Institute of the University of Pará (UFPA). Starting from the narratives picked up by the IFNOPAP project. We have developed studies on cultural terminology through oral narratives, shared by residents of eight cities in the Paraense Amazon.

In those histories, through their languages, we can observe their way of living, the habits, and the feelings and thoughts of the Amazonian people. These histories were translated into terminological vocabularies for each municipal district studied, in dictionary form. In that study, we saw the term as the conceptual unit that will help us compose the classification structure under the point of view of the culture. So, that cultural term, as part of the language, is an instrument to communicate the physical and cultural reality of the people that use it. Thus, when that man uses the speech or the writing for transmission of his culture, the terms become bearers of information on the culture.

In conceiving that term, the work approaches a theoretical discussion about the representation of the knowledge starting from terms of the popular narratives, inserted in different contexts, where the term is mediator between the understanding of the narrative and the information system.

If the term is a bearer of information concerning the culture of who uses it, then this cultural term reveals to us important aspects in the way of life, of the habits, of the faiths, as well as in the way to see the things. The RESNAPAP project understands this term as the symbolic representation of the cultural universe in which an individual is inserted, and, under the shelter of the linguistics, it can be seen about the lexical unit of the Amazonian man’s natural language (Oliveira; Rodrigues; Moysés, 2003b). This way, the objective of the work is to show the cultural identity, starting from the study, to elaborate the language of information that has as thematic universe to the Amazonian culture.
2. The Knowledge Representation and the Amazon Culture

The concept of knowledge representation can be seen starting from several definitions as they observe Davis et al. (1992a) (a) mechanism used to think about the world, instead of acting directly on it; (b) to answer the question “In what have I should think about the world?”; (c) a fragmented theory of reasoning, does it specify that inferences are valid and, which are recommended; (d) a middle of computation pragmatically efficient; and (e) a way of expression, that is, a language in which one can say things about the world. Thus, a representation is the language, through which we communicate and it can be classified in different ways, in agreement with the adopted theory.

For the General Theory of Terminology, man accomplishes the mental construction of an object and, when observing the reality that surrounds him, noticing individual objects, parts of those objects, and that he is inserted. Those particularities can be the color, the size, an action etc., that constitute the characteristic of the concept (Campos, 2001). Besides, Dahlberg (1978) points out that the definition is a resource used to minimize doubts in the use of the term, and more than that, makes possible, besides the fixed meaning of the concept, positions it in a system of concepts. Therefore, the Theory of the Concept of Dahlberg does not see the definition as an afterthought, but as a resource to establish the borders of the intention of the concept. For that theory, the characteristics are to understand the nature and the inclusion of the concept, besides they are responsible for the formation of the systemic conceptual structure. In that sense, they supply, at the same time, the skeleton, the bones and the tendons to systematize the knowledge (Campos, 2001; Tristão et al. 2004; Dahlberg, 1978).

The delimitation of the subject universe is inherent to a language of information, a thesaurus, because it systematizes concepts “of a subject area”. Thus, the thesauri are artificial systems of signs normalized with the objective of recovering, manually or automatically, the information that the user requests. In the definition of Tristão, et al. (2004), it is a group of concepts organized systematically for indexing and retrieval ends.

However, an interesting factor standing out is that, with the elaboration of the classification structure and, later, of the thesaurus, forming the language of information, we will also be contributing to the preservation of the cultural memory of the Paraense Amazonian, pointing out the cyclical dimension of the organization and detaching the knowledge in action, that, when organized and registered, starts to generate new knowledge.

The objective of a classification structure, just as of a thesaurus, is to sort things out, to represent the knowledge in an organized way, so that information can be retrieved for the needs of the user. The term “to represent” invokes a feeling of abstraction, or better, an abstraction of reality. A term is a sign in which man tried to represent an object in the closest way possible, given what he observed in empirical reality. So, after the perception, it designates synthetically the concept with its characteristic group, and it is stipulated to be called a term. Davis et al. (1992), they point out that the knowledge representation is a “substitute for that which represents” and they detach two fundamental aspects: first with relationship to the semantics of the representation and with relationship to the perfect fidelity because, for them, the representation needs an object that is the object in itself. They comment that the complexity of the real world is overwhelming in relation to the representational power of the concept.

With relationship to the characteristics used as beginning for division and comparison of the classificatory elements, working toward structuring the system starting from the formation of the classes and, inside of these, the arrays and the chains. Langridge (1977,
“Intensão” is related to the concept, differently of intention, that is related comfortable, desire. Array and chains they are terms given by Ranganathan to differentiate, in the formation of classes, horizontal series (arrays) — starting from a single characteristic of division — and vertical (chains) of concepts — in that each concept has more (descending) or the less (ascendancy) characteristics. Like this, Morin (2000, p. 24) affirms that all knowledge organization is accomplished in function of beginnings and rules and “it holds connection operations (conjunction, inclusion, implication) and of separation (differentiation, opposition, selection, exclusion)”, besides holding analysis and synthesis simultaneously.

For that, Ranganathan (1967), when outlining the steps of classification determines that the first step of the classificationist’s work in elaborating a classification structure is to map what he called Universe of Subjects. This activity has for function to define the level of extension of this universe. This map consists of deciding which subject areas that will be taken as base for the organization of the classificatory units and as this area it will be classified, which is a quite complex task for this author.

But, thinking about the dynamism of the universe of the knowledge, Ranganathan elaborates two canons as beginnings to think the formation of classes of concepts: the canon of exhaustivity and the canon of exclusiveness. In the canon exhaustivity, he establishes that the formed classes should be exhaustive, so that, if some new concept appears, it should be part of the structure, and this needs to have hospitality to contain it in an existent class. Already in the canon exclusiveness, Ranganathan establishes that the classes that form of the arrays should be mutually exclusive, in other words, no concept of the structure can belong it more than a class in the array. This way, the author does not accept polyhierarch as a tool (Campos, 2001).

Amazon culture is communicated in an environment where oral transmission prevails. As he observes Loureiro (1995, p. 55), that culture “contemplates in a predominant way the man’s relationship with the nature and it comes immersed in an atmosphere in that the imaginary privileges the aesthetic sense of that cultural reality”. That culture is observed as original and creative, because it reveals, it interprets and it creates its reality. Then, is important the text, the culture, and the geographical space in the study of the languages as knowledge representation (Oliveira, 2002). In the space of the forest and of the river they are the characters that are part of the Amazon culture. For example: CURUPIRA, that is a mythical character and his habitat is the forest in the space of his actions. The BOTO is a mythical character that inhabits the river and the ground. As Amazon fauna inhabits the river and as seductive man it inhabits the ground.

Still in the Amazon, the culture of the Indian stays in the narratives and it is outstanding its presence in the myths, habits, cookery and, consequently, in the mixed speech. As an example, the term JIRAU, comes from the tupi language (yula). It is a type of sink of rustic plenty wood usually placed in the window by the external part of the house. That architecture is used often by the riverside people that inhabit the stilts. As Oliveira observes (2003a, p. 247):

The oral narratives of the Amazonian, as a discursive process of memorization, contemplate and they portray the imaginary and the ideology, in an all complex one historical-partner-cultural of the Amazon people, together with the nature, the culture and the man between the river and the forest.

In that context, in the sense of preserving the cultural element, leaving of the man’s subjective vision, as part of that atmosphere and of its intrinsic relationship with the nature
that surrounds him. As Pombo observes (1988), the classifications of the beings, of the facts and of the events they constitute the stable points. They guide us in the world surrounds us and with that, establish habits, likeness and differences, help us recognize the places, the spaces and the events.

3. Methodology
The initial structure of the classification, for category, represents the knowledge in an organized way so that the information can be recovered by the user. In our methodology the delimitation of the subject universe is shown for the preservation of the memory and cultural identity of the Paraense Amazonian, because in this way it is registered and published, and as a subject universe it becomes information, restarting a continuous circle in the organization of the knowledge.

In the theory we studied the term, as a sign in which man tried to represent the most approximate characteristic of the object he observed in the reality that surrounds him. It is starting from this stage that we will know the characteristics of the concept, because they are the principal connection or separation elements, in the moment of formation of the classes. In that sense, they supply, at the same time, the skeleton, the bones and the tendons to systematize the knowledge. These characteristics are used as a beginning for division and comparison of the classificatory elements, working toward a structure of the system, starting from the formation of the classes and, inside of these, the arrays and the chains. Therefore, in this study we did not use a single method; but several, due to our technician-pragmatic character.

The content analysis prevailed with relationship to the analysis of the content of the concept, because there was the need to know it so that it was possible to insert its roots in the classification structure in agreement with its positioning adapted inside of the system. Therefore, so that it is possible to arrive at the elaboration of the classification structure for the representation of the knowledge, proposed in this study, some steps were indispensable and they preceded such activity. In the first step we read the picked up narratives and mapped in eight cities of the Paraense Amazon: Belém, Abaetetuba, Santarém, Castanhal, Bragança, Cametá, Altamira and Marajó’s Island. In the histories we identified the narrators as people that lived for a long time in those spaces. The purpose was to collect the cultural terms by the identification and analysis of these histories to register them for the terminological record.

The stage of the classification of the terms was only possible after the definition of the units that constitute the system, because starting from the definition of those units the characteristics of the same objects were elucidated. Later, it was necessary to identify the objects of same nature but related in the partitive and hierarchical relationships.

In what it concerns properly the classification, it was used to the categorical analysis, taxonomic method created “to satisfy the concerned collectors in introducing an order; according to certain criteria, in the apparent disorder” (Bardin, 1977, p. 37), as well as for refer association rules, of equivalence, or still of exclusion. Therefore, so that it roots possible to arrive to the elaboration of the classification structure proposed in the RESNAPAP project, some steps were indispensable and they preceded such activity.

The first accomplished step went to read the narratives popular orals, picked up in the eight municipal districts of the State of Pará. After identifying the classificatory unit, see as “cultural term” of the worked narratives, a significant step to structure the classification system, went to define the cultural terms. Soon after of the definition of the terms, they
took place with more precision the analysis and the selection of the cultural term, once, at this time, it was already obtained the domain of the extension of the concept. Finally, it took place the stage of the classification of those terms.

In the stage of the classification of the terms, initially, it grew up that Ranganathan (1967) is called Fundamental Categories. They were used by the author “to represent fundamental ideas”, in other words, including classes were created, that they could hold great number of subclasses and hierarchies, subordinated the larger classes. Created the Fundamental Categories, it felt beginning to the process of the ramifications and of the relationships of the concepts that would be inside the great classes.

It gave way, it happened the process of formation of the arrays and, later, the equivalence relationships were accomplished and of association. However, in some cases, that phenomenon became impracticable, because small number of concepts had connections or entails with different classes. In other words, a concept embraced more than one function in the Amazon culture as example, the term Miriti. It is so much an element of the flora as, raw material used in the Amazon craft.

<table>
<thead>
<tr>
<th>ART</th>
<th>BOTANICAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRAFT</td>
<td>BLOOMS</td>
</tr>
<tr>
<td>RAW MATERIAL</td>
<td>TREE</td>
</tr>
<tr>
<td>MIRITI</td>
<td>MIRITI (PALMS)</td>
</tr>
</tbody>
</table>

4. Analysis and Result

In the analysis, starting from the reading of the narratives of eight municipal districts of the Paraense Amazonian they were identified 1500 cultural terms. The registrations in terminological record created the terminological vocabularies of the spaces of the studied cities, in order to contribute in the understanding of the term, in the context of the narrative, forming a collection of six books. In the first book they are the picked up terms of the counted histories in Belém, Abaetetuba and Santarém. In the second, they are the removed terms of the histories told in Castanhal, in the third they are the removed terms of the histories told in Bragança, in the fourth, the terms of the histories of Altamira; in the fifth book with the removed terms of the histories of Cametá and, in the sixth book with the removed terms of the histories told in the Marajó’s island. After the books with the cultural terms and the definition, the content of the term is explicit in the context of the histories and to leave of those cultural terms, it went being created a classification structure, with the conceptual units, tends the following organization.

CS Categories of Subject
RT Relating Terms
ET Equivalent Terms

As a result is had the classification of the thematic universe with the terms in agreement with the certain categories, knowing the rich and fascinating culture of the Amazon people, full of secrets and mysteries, in the which the individuals produce a unique cultural environment. And, in that way observe the cultural identity, it can be said that is one of the few cultural forms that possess own characteristic, fundamentally because it is born in your own salience and it half feeds of your own one. Therefore it is denominated to that atmosphere of “Amazonian culture”. Besides, the work shows a theoretical discussion for
the understanding of the classificatory structure, in order to best to know the construction of a language that facilitates the use and the recovery in the system, denominated of language of information that is part of the study of the knowledge representation.

5. Final Considerations
This study of the knowledge representation, developed through the RESNAPAP project took us to a deeper knowledge on the rich and fascinating Amazonian culture. It also provided us a better understanding about the linguistics, the terminology and the theory of the classification.

It is important to say that in some cases, that phenomenon was difficult of being accomplished, because small number of concepts had connections or entails with different classes. In other words, that concept embraced more than a function in the Amazonian culture.

However, this work becomes important, because, with the elaboration of the classification for categories in the study of the language of information, the tendency is to publish and to preserve that culture in that a glance is focalized for that area, by virtue of it rich natural spring, fundamental for the scientific development, economic, medicinal, technological, among other aspects as the cultural of the inhabitants’ life with its habits, cookery, transportation means, house and myths.

References


Knowledge Organization in the Cross-cultural and Multicultural Society

Abstract
This paper explores the changing role of knowledge organization in the representation of information retrieval languages in the Cross-cultural and Multicultural Society. The paper studies the concept of Cross-cultural and Multicultural Society — the relationship between language and culture. Globalization and the Internet bring cultures, intellectual products and users closer to one another, but it shows their differences, too. With respect to KO the basic problem is: How can users from different backgrounds use the similar interfaces similar? The author introduces a Hungarian solution what maps the different topical terms, subjects, UDC codes on conceptual level.

1. Introduction
Nowadays a worldwide network infrastructure is rapidly moving information across national boundaries, but we often forget about the linguistic and cultural barriers, which make information retrieval more complicated. The solutions can be:

- the usage of multilingual thesauri
- multilingual subject headings
- the adaptation and usage of classification systems, which are not based on language, as UDC
- machine translation or machine-aided translation.

There are more and more emphases put in constructing a worldwide network of information. Information retrieval has become global, although a “global language” has not been developed yet. To some extent English can be called such, because it is used worldwide in different fields of research in development. Next to globalization another process can be observed as well: localization. Only those products which use the local language can be successful in a local market. Nowadays product designers use the local language and try to accommodate the special features of the local culture and customs. Local languages have become more and more important, although the network has stayed global. We can reach Norwegian, Korean or Dutch documents from Hungary, and Hungarian documents are available also from the furthest part of the world.

2. Language and culture

2.1. Language
European Union gives attention to evaluating policy measures for minority languages in Europe. There is a growing awareness of the importance of Europe’s linguistic diversity. “The yearly reports of European Centre for Minority apply in particular to minority language policies.” (Grin, 2000) The report in 2000 said:

Linguistic diversity has long been recognised as a defining characteristic of Europe, as an element conducive to the assertion of its identity, and as a condition for its democratic development; it is now also increasingly recognised as an asset enhancing creativity in all domains of social life, including social cohesion and economic performance. (Grin, 2000)

As we mentioned, English is one of the most important languages in global information retrieval, not the only one. Sometimes it seems like the only one used language in international and scientific communication. Researchers who investigate the development tendencies of the Internet say that nowadays the number of users who use English as a mother tongue is smaller that the non-English ones (Japanese, Chinese, Spanish). Other
research proved that the number of Non-English Web pages has overtaken the number of English Web pages.

That is why there is a growing need for cross-language retrieval systems, which help to overcome the languages barriers. A cross-language retrieval system is needed for those who can only search in one language, or for those who can speak more than one language, but his/her information retrieval is easier and more effective in one language.

Some typical cross-language retrieval tasks:
- search a monolingual collection in a language that the user cannot read
- retrieve information from multilingual collection using a query in a single language
- select images from a collection indexed with free text captions in an unfamiliar language
- locate documents in a multilingual collection of scanned page images.

2.2. Culture not only language
Culture is peculiar to Homo sapiens, together with material objects used as an integral part of this behavior. Thus, culture includes language, ideas, beliefs, customs, codes, institutions, tools, techniques, and works of art, rituals, and ceremonies, among other elements. There are 6000 languages in the world, which means almost 6000 different cultures, and means everyday problems in finding information in a databases or digital library originate from different cultures.

Hungary is a small country, but there are 13 recognized communities of ethnic minorities traditionally (Roma, German, Slovak, Croatian, Romanian, Serbian, Slovene, Armenian, Greek, Bulgarian, Polish, Ruthenian, and Ukrainian). Naturally there is much more nationality, but these raised minorities have special national ethnic culture in Hungary and own school systems, radio programs etc.

Human sociality is saturated with meaning (Holtgraves & Kashima, 2008, 73). The question of meaning making and meaning exchange — how people engage in meaningful thought and action in social contexts — is a central question of social cognition. (Kashima, 2000).

Definition of culture
Culture can be defined from many perspectives, according to the social science one is involved with. One of the well known is the definition by Kluckhohn.

He clearly distinguishes culture from the limited concepts of ordinary language, history and literature. The anthropological term designates those aspects of the total human environment, tangible and intangible, which have been created by men. A ‘culture’ refers to the distinctive way of life of a group of people, their complete ‘design for a living’. Culture seems to be the master concept of American anthropologists. For ethnologists, folklorists, anthropological linguists, archaeologists and social anthropologists, culture is always a point of departure or a point of reference if not invariably the point of emphasis. (Kluckhohn, 1951; Heidrich, 2002, 26)

Kroeber and Kluckhohn gave a very complex definition of culture:
Culture consists of patterns, explicit and implicit of and for behaviour acquired and transmitted by symbols, constituting the distinctive achievement of human groups, including their embodiment in artefacts; the essential core of culture consists of traditional (i.e., historically derived and selected) ideas and especially their attached values; culture systems may, on the one hand, be considered as products of action, on the other, as conditioning elements in a future action. (Kroeber and Kluckhohn, 1952)
2.3. Cross-cultural or cross-national

The differences between cultures are rare questions, usually the studies focus on the languages in the field of knowledge organization. Susanna Keränen studied this problem in her PhD dissertation (Keränen, 2002). She has examined equivalence of meaning, methods and dimension for measuring them. She presented many case studies in her practice under created an English and Finnish multilingual thesaurus. There are three problems of research in cross-cultural studies must be considered:

1. The term ‘culture’ has been used in so many ways that academics fail to arrive at a consensual definition;
2. The distinction between cultural and national boundaries is problematic with the consequence that ‘nation’ has been used as a synonym for ‘culture’;
3. The measurement of the impact of cultural attributes on organisational functioning is problematic due to the lack of definition clarity. (Clark, 1994)

Thus the other problem apart from the different aspects of cultural definitions is that ‘culture’ is often substituted for ‘nation’. The two terms are often used interchangeably. ‘Nation’ is invariably used as a synonym for ‘culture’ with the consequence that national distinctiveness is interpreted as cultural differences. Therefore, according to Bhagat and McQuaid (1982), “what are called cross-cultural differences are really only crossnational differences”.

Clark is correct in saying that it is an oversimplification to argue that cultural boundaries correspond to national (political) borders, since no nation is so pure that all its members share a common set of cultural factors. Any nation is a patchwork of different and unique subjective cultures. Thus, whereas two nations may share a common language, climate, political system and religion, differences in the mixture of their subjective cultures (subcultures?) will result in distinctive belief systems, norms, values and cognitive maps. A national culture therefore reflects the unique interaction between, and a combination of, a set of subjective cultures. There are many examples of these subjective (sub)cultures, which are clearly recognizable as regional cultures in Belgium, Canada, Germany, the UK, etc. (Heidrich, 2002, 27)

There are more levels and sources of Multiculturalism: personal age, geography, national affiliation, religious affiliations, language, tradition, historical elements, climate, economy, racial mix, political system, gender, modernization and many other intangibles. So if we are looking for any general knowledge organizational solution we should know the problem is very complex and a simple translation does not give a correct solution. Therefore the result on conceptual level can give sufficient accomplishment.

Features of culture:
- social product — it is based on mutual effect and interaction
- learned cognition and knowledge — socialization, system of cognition and norm
- individual — technical and intellectual appearance, what interpret the world by language
- different meaning — it is based on cultural experience

Propositions of culture:
- Culture is learned
- Culture derives from the biological, environmental, psychological, and historical components of human existence
- Culture is structured
- Culture is divided into aspects
- Culture is dynamic
- Culture is variable
Culture exhibits regularities that permit its analysis by scientific methods. Culture is the instrument whereby the individual adjusts to his total setting, and gains the means for creative expression. (Herskovits, 1940)

Situation of intercultural communication:
- Tourist
- Official visiting, scholars, delegation
- Immigrants and migrants
- Multinational workplace
- Minority, ethnical groups, subculture
- Internet users etc.
- People in the globalization

3. Equivalence in knowledge organization and digital environment

3.1. Equivalence
The basis of the problem is equivalence between worlds and concepts in multicultural and multilingual environment. We should define a threshold value, the equivalence level for KO solutions (multilingual thesauri or subject headings, the adaptation and usage of classification systems and machine translation or machine-aided translation). We should try to anticipate and decide what kind of topical term will most likely be used in our information retrieval system.

In International Organization for Standardization’s standards Documentation — Guidelines for the establishment and development of multilingual thesauri (ISO, 1985, p. 8) equivalence is divided into exact equivalence, inexact equivalence, partial equivalence, single-to-multiple equivalence and nonequivalence (Keränen, 2002, 797).

3.2. Digital environment
More and more digital libraries are designed and available all around the world. These libraries are produced in national languages, so they can be the target of cross-cultural and cross language information retrieval. The designer and content providers and users of the digital libraries can have different cultural backgrounds, so these libraries are cross-cultural in at least three levels. Cross cultural usability issues of digital libraries are an important area for social informatics. The misinterpretation of colors, form, symbols, metaphors, language and use are the most common cross-cultural design failures.

For number of years the focus of the research on digital libraries has been to solve the technical problems of digitalization and development. Usability has seemed a relatively minor concern. Although there were researchers and developers, who have investigated this cross-cultural issue considering the cultural aspects of designing of global information systems in a deeper way. This line of work can be grouped into four overlapping categories:
- cross-cultural usability studies
- internationalization of industrial products
- internationalization of software
- international web design.

Elke Ducker deal with cross-cultural usability, revealed in one of her researches the color preferences of students with different cultural backgrounds can differ dramatically. She pointed to the fact the usage of digital library and its metaphors are not unambiguous for
everybody. Computing metaphors have become an integral part of information systems design, yet they are deeply rooted in cultural practices.

Nowadays there are several projects in Europe which goal is to help and support the digitalization projects of the European countries, and to deal with the problem of cross-language retrieval in these digital libraries. Such projects are: Minerva and Michael, MACS (Multilingual Access to Subjects), Calimera, Madiera project (Multilingual Access to Data Infrastructures of the European Research Area) etc.

4. Multilingual thesauri
One of the solutions of helping the cross-language information retrieval can be the usage of multilingual thesauri. A multilingual thesaurus is more than just “putting together” several monolingual thesauri. A true multilingual thesaurus offers full conceptual and terminological inventories for each language represented; most importantly, it presents a fully developed thesaurus structure (i.e. all semantic relations of equivalence, hierarchy, and association/affinity) in each one of the languages of the thesaurus.

One of the main requirements of designing multilingual thesauri is that the different versions should be developed simultaneously. A multilingual thesaurus should not be a translation of an existing monolingual thesaurus. The practice although is not equivalent with the theory, because there are several multilingual thesauri which were developed by translation because of economic reasons.

Promising accomplishment is the Eurovoc thesaurus. The Eurovoc thesaurus is published in the official languages of the European Community (Bulgarian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Slovak, Slovene, Romanian, Spanish and Swedish) and one another language (Croatian). All these languages have equal status: each descriptor in one language necessarily matches a descriptor in each of the other languages. However, there is no equivalence between the non-descriptors in the various languages, as the richness of the vocabulary in each language varies from field to field (Eurovoc, 2008).

5. General Information Retrieval Language Dictionary — MÁTrlkSz
The usage of classification systems that are not based on languages can be another solution for the problem of cross-language information search, although there are several problems that have to be solved. Significant advantages are the conceptual level, the absence of linguistic signs among others, but the designers should observe, place and adopt the classification concepts to not only a foreign conceptual structure but establish connections with different subjects and topical terms.

The Mátriksz (General Information Retrieval Language Dictionary) project was finished by 2000 in the Széchényi National Library (Hungary), and there are some characteristics:

• It is the largest Hungarian Information Retrieval Language Dictionary on natural languages (about 28,000 lexical terms by 2002).
• It was made with the help of modern technology, by computer and it is used on electronic supporting, concerning informational network, Internet, too. Its software is RELEX.
• It is possible to develop and keep on the Internet dynamically.
• It can work with MARC format.
• It has very easy sentence construction and syntax.
• It is of high level, efficient, open, flexible, easily used, with rich vocabulary and a clear structure scheme.
• It has involved some of the existing Hungarian thesauri since 1990. Their numbers are over 20.
• It can be used by automatic classification and indexing systems, too.
• It can involve and the UDC index of the new Hungarian edition.
• The weakness of these thesauri was the few bibliographic records, which are really used in these systems. So the MÁTrIkSz project was a good opportunity for the KÖZTAURUSZ and the OSZK thesaurus. (Hajdu Barát, 2003)

The concept of project was: to give the possibility of the common search of the topical terms (subject, descriptors, UDC codes etc.) in some big libraries’ catalogues and databases with different structures and information retrieval methods.

The MÁTrIkSz is the co-operating system of the participant information seeking methods and databases. The KÖZTAURUSZ/OSZK thesaurus took a prominent role among them, but the participants are its equals. They are independent, and have the chance of common pursuit. There is a suitable user interface to search each participant dictionary and to utilize different topical terms, for example UDC terms, subjects, descriptors.

The MÁTrIkSz has its own information retrieval language dictionary, too, which is independent of the other bibliographical databases, but searches those systems. There is a common online index and its records have a local identification. The result is not only a bibliographic record, but there are contents of topical terms. The different classification systems and their bibliographical databases appear in a unified and homogeneous environment. Each database keeps their own descriptors, terms and if they have any similar expressions with different morphological forms, they make the KÖZTAURUSZ’s phrases appear in 750 field (equivalent) of the MARC record. Hits would come from all bibliographic databases, presenting diverse expressions of participant libraries without any analogy. Hits would come from only the own database (Bánki, 2002, p. 35).

The project assisted the new medium edition of UDC index. The adaptation of UDC MRF relating the KÖZTAURUSZ/OSZK thesaurus and the UDC index were built in the thesauri. The UDC codes became searchable in the MÁTrIkSz system, and they give the basis of searching in multicultural and multilingual environment, too, because the UDC codes and the descriptors are in conceptual level and the UDC independent of language.

6. Conclusion — Concentration to knowledge on the conceptual level
This paper aims at exploring the changes of knowledge organization in the role and possibility of representation of information retrieval languages in the Cross-cultural and Multicultural Society. Over the last decades there have been moves towards unification and standardization of bibliographic systems all over the world. This means the traditional classifications (for example UDC) and thesauri are now being stretched to cover cultural and linguistic artifacts and concepts would be quite different and broader from those originally intended.

References


Herskovits, M. J. (1940) *Man and His Works.*


Mixed Translation Models for the Dewey Decimal Classification (DDC) System

Abstract
This paper explores the feasibility of developing mixed translations of the Dewey Decimal Classification (DDC) system in countries/language groups where English enjoys wide use in academic and social discourse. A mixed translation uses existing DDC data in the vernacular plus additional data from the English-language full edition of the DDC to form a single mixed edition. Two approaches to mixed translations using Norwegian/English and Swedish/English DDC data are described, along with the design of a pilot study to evaluate use of a mixed translation as a classifier’s tool.

Introduction
Translations of the Dewey Decimal Classification (DDC) system are undertaken by authorized agencies around the world to produce localized representations of the DDC in which classes remain interoperable with the English-language edition on which the translation is based. Translations are localized with examples and terminology appropriate to the country/language group, and are often supplemented by interoperable expansions (e.g., a deeper representation of a geographic area than found in the English-language full edition). Beall (2003) discusses such expansions in detail, and provides several examples from the German and Vietnamese translations of the DDC.

We are exploring models for developing mixed translations of the Dewey Decimal Classification (DDC) system in countries/language groups where English enjoys wide use in academic and social discourse. A mixed translation uses existing DDC data in the vernacular plus English-language classes from the full edition to form a single edition. It is not a bilingual edition in the sense of parallel records in both languages; English-language classes will supplement classes in the vernacular. In some cases (e.g., geographic areas, history), developments in the vernacular will be in the form of interoperable expansions of their English-language equivalents (i.e., developments in which the English-language version is a logical abridgment of the expansion in the vernacular).

Why consider a mixed translation? A mixed translation might speed the translation process and make the translation easier to maintain. In general, the deepest levels of the mixed translation (with the exceptions noted previously) would be in English. The majority of updates to the DDC occur in classes subordinate to those found in the English-language abridged edition. Presumably, it would be easier to keep a mixed translation up-to-date by ingesting English-language records directly at deeper levels without going through the intermediate step of translation. The anticipated productivity gains in the development/maintenance of a mixed translation must be weighed against the usability of a mixed translation as a classifier’s tool and in end-user facing applications.

We are investigating the development of mixed translations through two case studies—a Norwegian-English mixed translation and a Swedish-English mixed translation. The two approaches share a number of common features. Both will use existing DDC data in the vernacular to form the base edition plus English-language data from the full edition of Dewey. Both will feature interoperable expansions in the vernacular in key ar-
Norwegian Model

There is a long history of use of the DDC in Norway dating back more than a century (Knutsen 2003). Current vernacular DDC content in Norwegian includes the 5th edition of Dewey’s Desimalklassifikasjon (Dewey 2002), a customized abridgment of DDC 21 commonly known as “DDK 5,” and the Norwegian translation of the DDC Summaries from DDC 22. The Norwegian model starts with the level of notation presented in DDK 5 as the framework for the vernacular content. The DDK 5 captions and notes will be revised to reflect the contents of equivalent classes in DDC 22. DDK 5 contains some longstanding alternative developments in geographic areas and history that will be addressed simultaneously in the English-language full edition and Norwegian content. Additionally, DDK 5 includes an optional arrangement for the law of Norway (based on one of the standard options presented in DDC 21); this will need special accommodation.

We plan a similar approach in the tables. DDK 5 currently contains Tables 1–7; Tables 3A–3C are collapsed into a single table similar to the one found in the abridged edition of the DDC. We will expand Tables 1–6 with English-language entries, but will include an interoperable version of the expansion for Norway in the vernacular content for Table 2. We will convert Table 3 into Tables 3A–3C.

The index to DDK 5 will serve as the basis for the Norwegian index, but full numbers will be included with segmentation marks to show logical abridgment. The English-language index will be included as a second index.

Swedish Model

Adoption of the DDC is currently under exploration in Sweden. Some vernacular DDC content already exists in Swedish—a mapping between the Swedish classification system (“SAB”) and DDC 21 schedule numbers (Gustavsson 2000), and a Swedish translation of the DDC Summaries from DDC 22. In a 2006 study, the National Library of Sweden explored the idea of a Swedish translation of the DDC (Svanberg 2006a, 2006b). There was general interest in a translation, but the abridged edition was rejected as too brief and the full edition as too detailed. It was also felt that a customized abridgment similar to DDK 5 would potentially create problems for interoperability and might be too time consuming to develop and maintain. A mixed Swedish-English translation arose as a possible solution.

The initial version will use the level of notation in the abridged edition of the DDC plus categories found in SAB as the guide for the categories to be translated into Swedish. Additional categories will be translated as needed based on literary warrant in Libris, the Swedish union catalog. English-language records will be added for the rest of the entries.

1. The DDC Summaries are available in Norwegian (Norsk) at: http://deweyresearch.oclc.org/ddcbrowser2/index_nor.html.
2. Table 7 was deleted in DDC 22 and replaced by notation derived directly from 001–999.
3. The DDC Summaries are available in Swedish (Svenska) at: http://deweyresearch.oclc.org/ddcbrowser2/index_swe.html.
The Swedish model will include full translations of Tables 1–6 into Swedish (with the exception of parts of Table 2, e.g., counties in the United States will remain in English). Eventually, additional vernacular content in the form of interoperable expansions in Table 2, history, and other areas of Swedish interest will be developed and added to the translation.

The mixed translation will include the full English-language index. Svenska ämnesord (the Swedish subject heading system) includes links to equivalent LC subject headings. We plan to investigate using this data to develop a mapping between Svenska ämnesord and Dewey. The Swedish subject headings could serve as entry vocabulary into the system, and as a possible basis for the development of the Swedish-language index. Swedish subject headings might not reflect the disciplinary focus of different Dewey numbers, but this could be addressed by including captions with the Dewey numbers under each Swedish subject heading.

**Methodology**

As a first step, we modeled two sets of number/caption pairs to illustrate the Norwegian and Swedish approaches. The first set is drawn from 370 Education. Figure 1 contains vernacular Norwegian data from DDK 5 (in bold) extended by English-language classes. Figure 2 contains vernacular Swedish data drawn from SAB / DDC 21 mappings (in bold) extended by English-language classes.

Note the Swedish-English mixed translation in figure 2 contains more classes in the vernacular than the Norwegian-English version in figure 1. This is because the Norwegian classes were pre-selected by the Norwegian translation team to form a specific level of abridgment in DDK 5, while the Swedish classes were derived from mappings between SAB and DDC 21 without pre-selection to form a cohesive Swedish version of the DDC.

The second set of number/caption pairs is drawn from 576 Genetics and evolution. In these examples, the vernacular content is at the same level in the Norwegian-English (fig. 3) and Swedish-English (fig. 4) versions. Since the footnote instruction applies to records in two languages, it is presented in the vernacular and English in both versions.

<table>
<thead>
<tr>
<th>Figure 1. Mixed Norwegian-English Translation in 370 Education (excerpt)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>370</strong></td>
</tr>
<tr>
<td><strong>370.1</strong></td>
</tr>
<tr>
<td><strong>370.11</strong></td>
</tr>
<tr>
<td><strong>370.111</strong></td>
</tr>
<tr>
<td><strong>370.112</strong></td>
</tr>
<tr>
<td><strong>370.113</strong></td>
</tr>
<tr>
<td><strong>370.114</strong></td>
</tr>
<tr>
<td><strong>370.115</strong></td>
</tr>
<tr>
<td><strong>370.116</strong></td>
</tr>
<tr>
<td><strong>370.1162</strong></td>
</tr>
<tr>
<td><strong>370.1163</strong></td>
</tr>
<tr>
<td><strong>370.117</strong></td>
</tr>
<tr>
<td><strong>370.1175</strong></td>
</tr>
<tr>
<td><strong>370.118</strong></td>
</tr>
<tr>
<td><strong>370.119</strong></td>
</tr>
</tbody>
</table>
Figure 2. Mixed Swedish-English Translation in 370 Education (excerpt)
370 Utbildning
370.1 Filosofi & teori, utbildning för bestämda mål, pedagogikens psykologi
370.11 Utbildning för bestämda mål
370.111 Grundläggande färdigheter
370.112 Fri uppostran
370.113 Yrkesutbildning
370.114 Moral, ethical, character education
370.115 Utbildning i samhällets grundläggande värderingar
370.116 Utbildning för internationell förståelse
370.1162 Internationellt studentutbyte
370.1163 Internationellt lärarutbyte
370.117 Interkulturell pedagogik och tvåspråkig utbildning
370.1175 Tvåspråkig utbildning
370.118 Education for creativity
370.119 Education for effective use of leisure

Figure 3. Mixed Norwegian-English Translation in 576 Genetics and evolution (excerpt)
576 Genetikk og evolusjon
576.5 Genetikk
576.52 Laws of genetics
576.53 Genetic makeup
576.54 Variation
576.542 Environmental factors (Mutagens)
576.544 Breeding patterns
576.549 Mutation
†Tilføy generell forminndeling slik den er modifisert i noten under 576 (Add standard subdivisions as instructed under 576)

Figure 4. Mixed Swedish-English Translation in 576 Genetics and evolution (excerpt)
576 Genetik och evolution
576.5 Genetik
576.52 Laws of genetics
576.53 Genetic makeup
576.54 Variation
576.542 Environmental factors (Mutagens)
576.544 Breeding patterns
576.549 Mutation
†Lägg till underindelningar enligt instruktioner under 576 (Add standard subdivisions as instructed under 576)

Discussion
Zeng and Chan (2004) note the challenges in establishing interoperability among knowledge organization systems when the source and target systems have different structures and characteristics. A mixed translation of the DDC will share a common conceptual structure at the class level, but there will be differences at the terminological level.

As we proceed with the study, we have identified a number of open issues for investigation related to content and representation. One key concern is the scattering of topics from including notes in Norwegian to subordinate English-language classes. When a Norwegian class is revised from an abridged version to one parallel to the version in the English-language full edition, certain topics currently appearing in Norwegian in including notes will be moved to an English-language class deeper in the hierarchy. The Norwe-
gian terms will still appear in the index, but the topics will no longer appear in Norwegian in the subordinate class itself.

Since topics are distributed throughout the DDC by disciplinary focus, a topic may be chosen for full translation in one place and retained in English in another. In figure 5, the record for 370.113 includes references in Norwegian to topics that appear in English in the record for 331.2592 Training.

**Figure 5.** References in Norwegian and English

370.113 Yrkesutdanning (fagutdanning) [Norwegian]

... Klassifiser fagopplæring i arbeidslivet, fagopplæring drevet av bedrifter i 331.259/2

... 331.259/2 Training [English]

... Class here interdisciplinary works on on-the-job vocational training, on vocational training provided by industry

Figure 5 also illustrates the use of a slash as a segmentation mark to show logical abridgment in notes and the number column. Smaller collections wishing to retain the abridged level of notation presented in DDK 5 can use the segmentation mark as a guide to logical abridgment of numbers. This information could also be used to create a wholly contained Norwegian version without the English-language records.

Some additional issues include handling of topics in scatter references, approaches to add tables and add instructions using data in two languages, and representation of relationships in the Relative Index in two languages. With respect to the last, our tentative plan is to provide two indexes—one in the vernacular, plus the complete English-language index. In an electronic environment, the two indexes can be browsed and searched simultaneously, but neither version will reflect the closely integrated structure between index terms and class records found in the indexes associated with the English-language full and abridged editions or with the localized translations of those editions. There will be instances of vernacular index terms pointing to concepts represented in records for which the language of description is English and vice versa. Perhaps this lack of symmetry at a terminological level is not troublesome as it might appear at first glance. In a discussion of multilingual thesauri, Hudon (1997) argues for acceptance of nonidentical and nonsymmetrical structures, and recommends that the number of descriptors in each linguistic version should be permitted to vary. The terminological content and structure of the Norwegian index will be at the very least equivalent to the version found in DDK 5 (and more extensive than the DDK 5 and the English-language versions in the areas with interoperable expansions). Additional Norwegian index terms could be added as needed to records in Norwegian or English.

**Pilot Study**

We are currently developing a full mixed Norwegian-English translation of an excerpt of the 370 Education schedule (complete with notes, add tables and instructions, and associated indexing) as the basis for a pilot usability study with Norwegian librarians in
second quarter 2008. Twenty Norwegian librarians were recruited in March 2008 to participate in the study. The pilot study group includes librarians from the national library and a library service agency, plus public, special, and university librarians. All are current users of the DDC. During a thirty-day period in second quarter 2008, the group will classify a pre-selected set of twenty documents in the field of education (split between works in Norwegian and works in English) using three tools: (1) the mixed translation schedule, (2) an excerpt from DDK 5, and (3) an excerpt from DDC 22 in English. All participants will complete an online questionnaire probing usability issues; follow-up semi-structured interviews will be held with study participants after analysis of the questionnaire responses.

We seek to learn more about the feasibility of a mixed translation as a classification tool by exploring such issues as:

- Should the level of notation in DDK 5 guide the level of classes in the vernacular in the mixed translation?
- Does the scattering of topics from including notes in abridged Norwegian classes to subordinate English-language classes affect the usability of the mixed translation?
- Should captions be presented in both languages at all levels of the translation?
- Should Norwegian index terms be used for some English-language classes in addition to English-language terms?
- Should other approaches be considered, e.g., a federated search of an abridged Norwegian database + the full English-language database?

If Norwegian librarians find the mixed-translation approach promising, the next step would be to proceed with a full Norwegian-English mixed translation. Whether to proceed with a Swedish translation of Dewey in any form is still under discussion in Sweden. The National Library of Sweden is considering a pilot study similar to the Norwegian one as a step in informing the decision.

Conclusion

We have undertaken a study of mixed translations of the DDC as an approach to streamlining the development and maintenance of a translation in countries / language groups where English enjoys wide use in academic and social discourse. The initial study is focused on the usability of the mixed translation as a classifier’s tool. We have not studied the implications of mixed translations for end-user facing applications beyond the notation itself. This needs to be considered in future work.

Acknowledgments

The authors are grateful for the advice and comments of the Dewey editorial team, Pia Leth (National Library of Sweden), and colleagues at the National Library of Norway. DDC, Dewey, and Dewey Decimal Classification are registered trademarks of OCLC Online Computer Library Center, Inc.

References


Discovery and Access Systems for Websites and Cultural Heritage Sites
Reconsidering the Practical Application of Facets

Abstract
Facets are an increasingly common feature of contemporary access and discovery systems. These intuitively adaptable structures seem well suited for application in information architecture and the practice of knowledge management (La Barre, 2006). As browsing and searching devices, facets function equally well on e-commerce sites, digital museum portals, and online library catalogs. This paper argues that clearly articulated principles for facets and facet analysis must draw examples from current practice while building upon heritage principles in order to scaffold the development of robust and fully faceted information infrastructures.

Introduction
Reference to the theoretical structures that underpin the creation and use of facets is often absent from current discourse about ‘faceted’ resource discovery and access systems. This omission obscures the fact that facets, as an information retrieval construct, arise from practice, from observation, and from use. The roots of facet theory extend deeply into early implementations of document retrieval systems. One of the earliest faceted systems, the Universal Decimal Classification, was created by the documentalists Otlet and La Fontaine in order to provide practical access to ‘facts’ in the form of bits of information pasted onto cards in a card catalog with the hope of creating an immense synoptic outline of the world’s knowledge (Rayward, 1997). Later, Ranganathan tested facets as an alternative to the hierarchical straightjacket of traditional classification systems, by implementing a faceted access system, the Colon Classification, in an academic library. After reclassing the library, Ranganathan and his staff observed patrons, tallied requests, and corrected issues before fully formulating the principles of facet analysis and detailing the principles of faceted classification. In Ranganathan’s words (1971), “After a long experience is gained with an improvised aid, a theory is developed in order to understand the aid deeply and to systematize, improve, refine and develop it.”

Contemporary support for faceted approaches appears in various forms. Faceted access to folksonomies has been proposed as a way to tame unwieldy word clouds (Quintarelli, Resmini, Rosati, 2007). Facets and facet analysis are also mentioned as valuable aspects of two critical discovery and access standards, NISO Z39.19, Guidelines for the Construction, Format and Management of Monolingual Controlled Vocabularies (2005, pp. 14–15) and the current draft of Functional Requirements and Numbering of Authority Records (FRANAR) (2007, p. 23).

Given these practical aspects of and current support for systems that draw from the heritage of facet theory, difficulties and issues that arise from current practice may prove to be a fruitful area for extending or strengthening traditional principles. Without such knowledge it can be difficult to overlook the fact that facets don’t exist in a vacuum. Facet analysis must be anchored in anticipated use and in the domain of application. North American use of facets has been largely confined to small domains to provide access to objects (not concepts) in commercial or business applications. Many current implementations seem to be unaware of the existence of previous research and implementations dating to the 1950s (Atherton, 1961; Cleverdon, and Keen, 1966). European understanding
and application of facets is more nuanced, owing to greater experience with facets and their use, by subject specialists for organizing and creating access to special collections in business, government and special collections (Broughton, 2006).

In North America, the term facet is frequently used as a new way to describe pre-existing data fields in legacy systems such as bibliographic records and the back-end relational databases on many e-commerce sites. In such a case it is not often clear that facet analysis, the critical first step in the construction of faceted access, has been conducted. Allowing the underlying data structures to fully determine all facets in use in a facet ed navigation system may serve to inhibit the implementation of full featured systems. Other less obvious facets, or those that are not hardwired into an existing system, may emerge from conducting rigorous facet analysis and thereby permit iterative development of more powerful and dynamic systems even as they are grounded in facet theory.

Understanding and use of facets in North America began to change drastically in early 2006 when North Carolina State University (NCSU) became the first library to implement the faceted or ‘guided navigation’ system of Endeca software in their library catalog (OPAC). Many other libraries followed suit, with the assistance of commercial vendors and the open source community. This heralds a re-entry into a time of exploration, exploitation and testing of the power of facets and various other approaches to enhancing the search and discovery experience. The NCSU catalog itself is not static as functionality is still being added in a stepwise fashion, but remains one of the most complete demonstrations of faceted access currently available. Many of the library and museum interface and software products that use facets are in beta versions, many are open source. A large number of new implementations are planned during the latter half of 2008.

Methodology
In order to scaffold the development of fully faceted infrastructures, clearly articulated for facets and facet analysis must be anchored in examples drawn from current practice. In order to assess current practice, this exploration revisited a stratified random sample of 200 websites from an earlier examination of the use of facets on the web (La Barre, 2006). These sites were revisited in 2008, and changes in facet use on the remaining sites (n=170) were recorded using a content and facet analytical protocol tested during the earlier study (Aitchison, Gilchrist and Bawden, 2002; Atherton, 1961; Broughton, 2006; La Barre, 2006). ‘Faceted’ websites from this sample were compared with a small set of newly implemented ‘faceted’ interfaces in libraries and museums (n=9).

The original (2005) sample of 200 sites was drawn from the Open Directory Project’s (ODP) DMOZ volunteer built website directory (dmoz.org). At that time, DMOZ consisted of seventeen main categories and contained links to over four million sites. This seemed to offer a set of current websites that could provide a snapshot of common design practices. During the website survey phase of data collection in 2005, 200 sites were randomly selected from four DMOZ categories: (for-profit): Shopping, Business and (not for profit) Society and Reference in order to maximize comparability. Each randomly selected site was subject to a series of validation procedures to ensure currency, and comparability.

The 2008 survey revisited the sites which remained online and freely accessible (n=170). These sites formed a base set for comparison with a set of six online library catalogs and three museum interface prototypes that self-identify as using facets for a total of 179 sites in this sample (see Table 2). Five of these implementations use open source
software, three in libraries and two in museums, the rest are commercial software applications.

Data for each survey (2005 and 2008) was collected over a two day period to reduce variability of results. Content analysis for site components including search and browse areas, site maps and navigational structures was conducted. A wireframe was created for each site; the basic and/or advanced search mechanisms were tested using the query ‘search’, the sitemap and first three levels of the website (including any search result screens resulting were printed out for later reference. Each wireframe was coded for the features described above and facets were identified.

The 2005 study indicated that web designers often seem to be following Rosenfeld and Morville’s (2002, p. 170) recommendation to integrate search and navigation (or browsing) systems on websites in order to allow users to jump easily between the two in order to conduct searches which lead to browsing, or searches which retrieve categories and documents. Thirty-six percent of the sites used integrated search and browse structures. “Instead of forcing one way to view the items, Faceted Navigation allows users to view the items in any way they want. At the same time they learn how the items are structured so that they may consider other search strategies in the future” (van Welie, 2004). Integrated search was also observed on each of the nine library and museum applications included in the 2008 study.

Discussion

Most of the sample sites were designed to provide access to concrete items, such as consumer products, documents or artifacts, rather than subjects or ideas. Facets on all of the sites were captured by referring to Vickery’s steps in facet analysis (1960, 1966): (1) Define the subject field by determining what entities are useful to the user. (2) Formulate facets by examining material that expresses the needs of the user group to obtain candidate terms. (3) Structure each facet by creating a hierarchical order and to identify gaps and overlapping concepts (for a complete description of the facet analytical protocol see La Barre, 2006).

Surprisingly, only 30, or 15% of the websites in the 2005 sample were offline, or had changed to subscription sites. This resulted in the loss of six previously identified faceted sites. Five of the offline sites were faceted. Another previously identified faceted website dropped from the sample when one of the shopping sites abandoned the use of facets in a redesign and removal of the browsing and advanced search features. Over half of the remaining 170 sites had been updated in some way, and most had changed their navigation or search features, adding facets in the process. Business sites showed the greatest levels of updates to the sitemap, search and navigation components, while the Shopping sites showed the least number of changes to these components.

The remaining sites showed an increase in the use of facets on sites that had not previously used them. Shopping (23.4% increase) and Business (22.3% increase) had the highest increases in use, followed by Society (18.4%) and Reference (14.6%).

Sites with newly added facets implement them as features of browsing or searching interfaces. Reference sites used faceted browsing interfaces. New facet use on Society sites concentrated on search. On Business and Shopping sites, new facet use was equally split among search and browsing areas. No sites from the 2005 sample showed new facet implementations in both areas.
Facet use in libraries and museums

The library catalogs chosen for inclusion in this website survey have ardent supporters and implementers who have vigorously promoted, and documented the implementation of each catalog. Table 1 contains a listing of the sites in this subset. Each of these implementations demonstrates improvements over the traditional e-commerce approach to facets. Each also represents current best practice in providing faceted access and discovery. Three major vendor implementations were selected which use software from Endeca, Exlibris and OCLC. Multiple implementations of each of these products are ongoing at various libraries across the world. The three library open source products selected also represent exemplary approaches that are being closely watched by system designers, and include the Fac-Bac-OPAC, LibLime’s Koha Zoom, and the University of Virginia’s Blacklight project. The museum prototype implementations are drawn from major projects including the Flamenco project, the OpenCollection project’s newest Access module implementation, and an older digital library implementation at Indiana University. Each of these museum projects has provided leadership and guidance for members of the digital museum community.

In comparing each of the implementations alone, it would appear that either the metadata in bibliographic records and museum records is far more complex than that on the e-commerce sites, or e-commerce metadata is severely underutilized. Important differences emerge between e-commerce applications and the library and museum implementations when location of facet use and numbers of facets in use on each site are compared. These findings can only be indicative, not fully informative because of the small sample size. Table 2 describes facet use by site function. Facets on the six library sites are used in both search and browsing components, while two of the museum sites use facets for browsing only. This is rarely the case in the DMOZ sites, as only one of the eighteen Business sites, and three of the twenty Society sites use facets in both areas. Average numbers of facets in use are revealing and range from an average low of 1.6 facets for search refinements use on the five Reference sites, to an average high of 4.4 in the seven browsing areas of the Business sites. Library use of facets in search interfaces averaged 10 per site, and 5.8 per site in browsing areas. Museum use averaged 7.3 facets per site and 7 for browsing areas.

Table 3 demonstrates facet use on the sample sites. As expected most sites used unique labels or facets. Terms below reflect those in actual use on the sample sites. While some of these facets may seem to be applicable only in libraries, some such as brand or price clearly seem to belong to the world of e-commerce. Facets such as topic/subject, format, language, author, location, date and year are used across all site categories and components.

Many questions remain for future work in this area. Might it be possible for library implementations to more completely harness the richness of the data in bibliographic records, while learning lessons from e-commerce implementations? Will the use of integrated search and browsing components be as useful to library users as they seem to be for e-commerce site customers? Are the relatively low numbers of facets used in the six implementations reflective of interface or software limitations? Are there generic facets that are useful across domain that are not being harnessed in websites? Are some facets more useful as searching or browsing aids? Might facet analysis suggest facets not currently being used, but useful in given domains?
Table 1. Sample of ‘faceted’ Library and Museum implementations (O/S denotes open source)

<table>
<thead>
<tr>
<th>Library applications</th>
<th>Museum applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endeca</td>
<td>North Carolina State University</td>
</tr>
<tr>
<td>Exlibris Primo</td>
<td>Vanderbilt University</td>
</tr>
<tr>
<td>OCLC Worldcat</td>
<td>U. Washington</td>
</tr>
<tr>
<td>Fac-Bac-OPAC (O/S)</td>
<td>Paul Smith’s College</td>
</tr>
<tr>
<td>LibLime Koha Zoom (O/S)</td>
<td>Nelsonville Public Library</td>
</tr>
<tr>
<td>UVA Blacklight (O/S)</td>
<td>University of Virginia</td>
</tr>
<tr>
<td>Flamenco/Thinker.org (O/S)</td>
<td>University of California, Berkeley</td>
</tr>
<tr>
<td>Open Collection Access/Coney Island Historical Project (O/S)</td>
<td>New York City Department of Cultural Affairs and New York City Department of Youth and Community Development</td>
</tr>
<tr>
<td>Charles Cushman Photographs</td>
<td>Indiana University</td>
</tr>
</tbody>
</table>

Table 2. Average facets in use by component and category. n=sites with facets in use. Parentheses indicate number of sites using each component per category

<table>
<thead>
<tr>
<th>Searching</th>
<th>Browsing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference (5)</td>
<td>(1) 9 (5) 1.6</td>
</tr>
<tr>
<td>Business (18)</td>
<td>(12) 3.5 (7) 4.4</td>
</tr>
<tr>
<td>Shopping (20)</td>
<td>(13) 3.48 (10) 3.4</td>
</tr>
<tr>
<td>Society (7)</td>
<td>(5) 3 (2) 3</td>
</tr>
<tr>
<td>Library (6)</td>
<td>(6) 10 (6) 5.8</td>
</tr>
<tr>
<td>Museum (3)</td>
<td>(1) 8 (3) 7.3</td>
</tr>
</tbody>
</table>

Table 3. Most commonly used facets in the 179 sample sites

| topic/subject | 29 |
| location | 19 |
| author | 10 |
| date/year | 8 |
| country/region | 7 |
| content | 5 |
| when | 4 |
| category | 23 |
| brand | 13 |
| price | 10 |
| title | 8 |
| genre | 6 |
| keyword | 4 |
| form(at) | 19 |
| language | 11 |
| type | 9 |
| class number | 7 |
| library | 6 |
| what | 4 |
| keyword(s) | 5 |

Conclusions

It can be difficult for an information architect or a knowledge management specialist to convince clients that time invested in detailed facet analysis is worth the cost (La Barre, 2006). This has resulted in a number of partial implementations of faceted systems. Existing and commonly referenced faceted systems have limitations that may echo the limitations of the systems they seek to replace. Too often these may evince no engagement with facet theory, or engage the principles superficially because they were developed in the time-bounded business culture that necessitates rapid return on investment (ROI). In this climate, without knowledgeable intervention, wide scale adoption of faceted systems, such as we are now seeing in museum and library catalogs, may be fatally flawed from the outset. A re-examination of the extensive failure analysis and use studies such as those in the Cranfield studies, as well as principled application of facet analysis that were common features of early faceted information retrieval experimentations may well be useful to those considering or implementing faceted access systems. Rapid ROI is the order of the day in the economic climate that faces libraries and businesses. Will this prove to be
the acid test of facets? Many practitioners have requested that the residents of the academy fill these gaps with conscientious use and user testing in order to provide support for their anecdotal understanding of the power of facets and facet analysis. Concerted, well designed evaluations are rarely implemented in the business world. It is vital to reconsider the intriguing symbiosis between theory and practice during this period of rapid deployment and testing of the information infrastructure for the twenty-first century. This research seeks to continue a long tradition of drawing from contemporary practice in order to ensure that facets are understood as practical extensions of theory, which must be grounded in practice and not as dry, sterile devices devoid of connection to practice.

References


On the Relation Between Qualitative Digitization and Library Institutional Identity

Abstract
The paper highlights and discusses concepts and practices of national library digitization. Two conceptual models are suggested in order to help strengthen scholarly analyses of digitization practices within libraries: the distinction between quantitative and qualitative digitization, on the one hand, and the prism metaphor for understanding the nature of qualitative digitization, on the other. Qualitative digitization, understood as a document representational practice, is defined as a knowledge organization practice. These concepts and models are then put in relation to the issue of national libraries and institutional identity. By combining research within the fields of KO, digitization and institutional identity, the paper points to a way of addressing empirical issues in all three fields of research. Special focus is on the very definition of qualitative digitization as a KO practice and the practice of selecting and digitizing documents suitable for the development of institutional identity within libraries.

Aim
The aim of this paper is to discuss and rethink concepts of digitization and knowledge organization (KO) practices in relation to cultural heritage digitization and library identity. By referring to practices in Scandinavian national libraries and institutional identity theory, claims are made that document digitization and markup should be considered as KO practices and that they play an important part in shaping cultural heritage and the very identity of national libraries as institutions. Rather than producing answers, our aim is to provide a basis for further questions and thus create a platform for future research. It is our belief that placing KO practices in new contexts and in relation to current trends in the ALM (Archives, Libraries, Museums) sector, we may contribute to a relevant scholarly discussion in Library and Information Science (LIS).

The paper starts off by discussing different forms of digitization as practices of KO. Turning to research on institutional identity relevant for the ALM sector, conclusive suggestions are made on how to develop research on the relation between qualitative digitization and library identity as part of the theoretical development of KO.

Digitization as Knowledge Organization practice
Let us begin by making two clarifications. First, we make a distinction between quantitative and qualitative digitization. The former, sometimes referred to as “mass digitization” and exemplified by the Google Books, the Open Content Alliance or the European i2010 efforts, aims to digitize massive amounts of documents (thus an all-inclusive strategy) using automated means (Coyle 2006). For pragmatic reasons, mass digitization has to disambiguate the transmission phases, minimize interpretation and flatten out the digitization process into a two-dimensional linear affair. Qualitative (or critical) digitization (henceforth shortened as QD) on the other hand involves deep text encoding, critical image and text editing, and rich information assignment to critically selected documents, a more exclusive strategy. Secondly, when we are referring below to text encoding, we are talking about descriptive strategies using element-rich XML-compliant schemes such as the TEI (Text Encoding Initiative), where the contents and structures of a document are assigned descriptive labels, as opposed to procedural and presentational strategies, where
encoding is used primarily for visual effects (Renear 2004). The TEI scheme is appropriately employed within qualitative digitization projects, while quantitative digitization uses simpler text encoding, if at all, merely for presentational (and in this context more trivial) purposes.

Cultural heritage digitization in the ALM sector has relevance for, and is being studied within, several areas within LIS (e.g. collection development, digital libraries, book history, open access and publishing studies). Knowledge organization (KO) is arguably the LIS area where digitization, and in particular QD, proves to be most fruitful for testing and development. For instance, digitized document collections and the way they are internally organized offer considerable challenges to fundamental bibliographic concepts and models (Dahlström 2006, ch. 2; Renear et al. 2003). Further, digitization practices prove to be intimately engaged in KO practices — to the extent that we argue for considering QD as partly a KO practice. A major area of KO can be called document representation (DR), exemplified by metadata production, bibliographic practices and descriptive cataloguing.1 In DR, documents and their components are analyzed, labels and other metatexts are assigned to them, and textual parts of documents are transmitted into a new, target document, a representation. This representation can range from being quite brief, as a catalogue post, to being a more or less full simulation of the departure document, where specific codes are assigned to the document sections at various granularity levels (Dahlström 2004, 21). QD using detailed descriptive encoding is another instance of such DR.

QD is connected to cataloguing and metadata in three ways. Firstly, proactive as well as reactive digitization (Lee 1999) often necessitates a new and separate cataloguing of materials that have so far not been catalogued at all, usually because the materials are component parts of other documents, e.g. single sheets or graphic objects. Secondly, sophisticated descriptive text encoding demands a relatively exhaustive metadata account of both the original and the digital document (in TEI this is performed using a so-called TEI Header). Thirdly, the very application of a markup scheme and its elements means structuring, labelling and transmitting (parts) of documents in a way much similar to other DR. Such DR practices within QD are conditioned by norms and strategies that are either described in policy principles (“best practices” and guides) or formalized in standardized schemes.

**Qualitative Digitization as prism — a tentative metaphor**

Where quantitative digitization implies a simple linear transmission model (simply having the book scanned and putting the resulting image files on the web), qualitative digitization involves intellectual, interpretative and multi-faceted measures at several phases of the digitization chain. Such critical, interpretative measures have for long been recognized by textual critics and scholarly editors as the basis of their practices, and QD, applied rigorously, indeed becomes an application of scholarly, critical editing. Perhaps an even more suitable term would therefore be critical digitization.

The *input* phase of QD includes e.g. policy making/adaptation and the identification, selection, collation, analysis, description, cataloguing and classification of source documents. The *preparation* phase involves such tasks as image and text capture, transcription/OCR, conversion, image enhancement, intellectual text encoding, database construction, annotations, as well as metadata work with the digital material. The *output* phase,

---

1. With subject representation and IR as the two other major KO sections.
finally, has to do with activities such as manual or automatic transformation (e.g. using XSLT), versioning, layering, linking, publication, availability, archiving, optimization, added values, feedback, marketing, evaluation, and the important questions of use and re-use. What comes out in the end is not a simple clone of the analogue source, but an enriched, interpreted, changed and new object that to a certain degree reflects but also comments on the source. A user approaching a digitized collection thus faces an interpreted and argumentative material that answers the particular questions the digitizing agent has asked during the digitization process. As noted by Rockwell (2003, 215) however, researching a digitized material means being able to subject it to new and unanticipated questions — a scholarly ideal that is conditioned and, arguably, hampered with each intervening interpretation.

In this sense, QD can be understood as a prism. Prisms reflect, deflect, or transmit light, and are used in e.g. cameras to split light into a spectrum of colours and with the aid of mirrors create a representation of an object by producing a simulation of it. Similarly, QD splits a more or less cohesive analogue object into a spectrum of files that vary in category, function, extension and level. In QD, one would e.g. end up with packets of files such as raw texts with stand-off markup, or text files with embedded marked up (.xml or .tei), document type definitions — i.e. markup rules for a genre or document type (.dtd), style sheets (.css), transformation instructions (.xsl), facsimile source files (.tiff), image delivery files (.gif or .jpg), metadata (in RDF or teiheaders), databases etc. These different spectrum files may then be edited separately, combined and re-combined at will, each constellation bringing a new representation of the source as an altered simulation.

The agent performing QD, such as a library, faces most of these possibilities at various phases in the process, and has to make choices of what to select and what to leave behind, what to enrich intellectually, and what to make available to which audience in what form for what purposes. The library’s implementation of DR/KO practices such as versioning, bibliographic relations, text encoding, document organization and distribution is instrumental in that process, and conditions the possible ways the material selected can be made available, such as in what formats and various outputs, and what possible uses and re-uses can be made of the material. In so doing, the digitization design legitimises particular kinds and levels of materials at the expense of others and addresses particular needs and audiences at the expense of others — a question of symbolic power.

The library’s decisions and overall QD strategy for selection, preparation as well as availability has significant bearings on the overall shape of the accumulated digitized library collections. Qualitative is more likely than quantitative digitization to make the digitizing library attentive to the intellectual, interpretative features of cultural heritage digitization. Intellectual DR practices such as text encoding and other forms of descriptive metatext labelling are important in shaping the prism trait of QD, the emerging pattern of the national digital library holdings, and to some extent therefore the cultural identity of a national library.

National libraries and institutional identity

Issues on institutional practice and identity have been dealt with in several disciplines, most prominently perhaps in political science and sociology (March & Olsen 1989, Peters 1999). Within LIS it has been brought up by Frohmann (2004), and identities of libraries have been discussed by Audunson (1999), Goodrum (1980), and Hansson (2006). Most of these studies take their departure in the relation between norms, rules and associated
practices in defined types of libraries such as public libraries or national libraries. The identity of an institution manifests itself through daily practices and decided policies. The definition of institutional identity however goes beyond this. A national library forms its identity in relation to the idea of the national state which it is supposed to reflect in terms of collection development, for instance through legal deposit regulation, but also through the relation it has to the rest of the library sector. National identity is mirrored in the identity of institutions. In Europe there has, by tradition, been a sharp division between e.g. public libraries and the national library, while in many younger national states, such as many of those in Africa, we can see a much closer integration of the different parts of the library sector (Hansson and Kawalya 2007).

KO practices such as document representation and subject analysis are significant tools in library practice that help define this idea, the set of values underlying the institution (Andersen 2004; Hansson 1999). Kjellman (2006) shows how institutional practices directly influence the practice of KO in her domain analytical study of the picture collection of the National Library of Sweden. Reversely, she concludes, the practice of KO constitutes a major influence on the general institutional practice of the National Library.

Over time we can see different practices emerge and perhaps the most prominent one in later years is that of digitization of various parts of the collections or specific documents. This practice contributes significantly to the totality of identity-shaping practices and, as it is relatively new, to the development of the identity of library institutions in contemporary society. The shaping of an identity is on the one hand based on practices that come out of the formulated mission of the library, but it is also a result of intentional decision-making regarding how the library wants (or expects) to be perceived by the outside world. Depending on the points of departure in individual cases, there are thus many ways of letting the process of digitizing collections contribute to the shaping of a developed identity of the institution. Two examples of the variation of decision-making and digitization practices are the national libraries of Norway and Sweden.

In Norway, the National Library is pursuing the practice of digitizing, more or less, their full collection. The library today has over one million documents digitized. This practice is based upon quantitative digitization, which, as suggested earlier, hinges to a lesser degree on the intellectual processes of KO than does qualitative digitization.2 The National Library of Sweden, on the other hand, is presently engaged in a high degree of selection and refinement in the individual expressions of the digitized document, produced in small-scale projects. This obviously means applying qualitative digitization, where unique and historically high profile documents are chosen from the collections and subjected to detailed text encoding and analysis.3 In all aspects, the digitizing practice of the National Library of Sweden follows the intellectual practice of interpretation that is so characteristic of qualitative document representation. As such it is an obvious renewal of traditional KO practices. Why is this so important to stress? It is so because of the impact KO has on the identity of libraries as social institutions.

2. It should be noted that the National Library of Norway also works with more limited, qualitative and critical digitizing projects in parallel to its mass digitizing policy. An example of this is the collected manuscripts and letters of Henrik Ibsen, http://www.ibsen.uio.no.

3. Two examples of such qualitative digitization projects of the National Library of Sweden concern the Codex Gigas (http://www.kb.se/codex-gigas/eng/) and the book Suecia antiqua et hodierna (http://www.kb.se/samlingarna/digitala/suecia/) — the latter being, as a case in point, a classic and monumental expression of the formation of Swedish national identity during the 17th century.
From this point of view, it is clear that these two approaches both contribute, but in very different ways. Digitizing practices contain more or less conscious processes of interpretation. When ALM institutions engage in mass digitizing, they are likely to make use of a simplified media model which trivializes cohesive documents and their transmission, treating the latter as mere cloning. Kjellman (2006, 239 f.) further observes that whereas museums tend to consciously discriminate and select, the "totality ambition" of libraries when collecting materials results in notions of neutrality, which tend to hide the institution’s discriminating mechanism. Such a tendency is obviously fuelled by quantitative digitization. Qualitative digitization on the other hand is more likely to emphasize the interpretative aspects of the KO principles underlying the process, as described by the prism metaphor. The choices made are conscious, and the result usually defines not only the actual process in a narrow technical sense, but in an institutional and socially relevant way as well.

Cultural heritage (CH) digitization in libraries has two main objectives: preserving the materials and making them available to current and future audiences, and thus also to support use and re-use of the materials within education and research. To support such objectives, the issues of what material is being selected as input and what kind of material is being made accessible as output — and how — become crucial factors. The selected documents fill a symbolic function as representatives for the contents of the whole library. So does the way in which they are bibliographically organized. The selection of documents to be exhibited in digitized forms is crucial in forming the identity of the institution as accessed through the web. KO practices are among those that form the basis of this identity-shaping selection of cohesive documents to be fragmentized in the elaborate digitizing process. The way in which this is done in practical library development, however, needs to be further analysed.

Conclusion
The relation between digitization, KO practice and the development of institutional identity within ALM institutions has not been subjected to analysis in any major way. At the same time, digitizing projects, both qualitative and quantitative, are put forward as crucial to the libraries, archives and museums involved. Digitization is seen as a way for these institutions to keep up with the Zeitgeist and often provides a major argument for the relevance of them in contemporary society, both on a local and regional level. In our coming research efforts in this field, we will focus on the selection and accessibility phases. More precisely, we will address questions such as:

• How do the publication formats include or exclude audiences, and how do the KO methods applied support use and re-use of the digitized material?

Relevant problems to study include text or image capture strategies; granularity and versioning; text encoding and transformation; restricted versus “open source” access; scalability, modularity and longevity (cf. Besser 2004; Bodard and Garcés 2008; Dahlström 2008; and Seaman 2003). Those studies will require analyses of what kinds of material have been selected and why, at what audiences they are targeted, and what possible implications that might suggest in relation to the identity discussion implied above. In other words, when focussing on national libraries and on the issues of selection and availabil-

4. Individual examples of such analyses can of course be found. See for instance Dalbello 2004.
ity as manifest practices in relation to an identity which conjuncts with the national state, several issues of concern emerge:

- In what way do the digitized and organized material manifest issues of identity in terms of functionality and availability for various kinds of users?

The issues raised in this paper should be seen as part of a discussion of the impact of new practices in librarianship. Practices based on KO principles are the very core of this development. Using the distinction between quantitative and qualitative digitization along with the prism metaphor for combining relevant discussions of institutional identity, practical digitization and KO principles, provides us with an opportunity to raise questions concerning these socially and institutionally highly relevant relations in a more informed way.

Acknowledgement

We would like to express our gratitude to the two anonymous reviewers of this paper. Their comments helped to improve its content.

References


Amelia Abreu — University of Washington

“Every Bit Informs Another”
Framework Analysis for Descriptive Practice and Linked Information

Abstract
The independent traditions of description in bibliographic and archival environments are rich and continually evolving. Recognizing this, how can Libraries, Archives and Museums seek convergence in describing materials on the web? In order to seek better description for materials and cross-institutional alignment, we can first re-conceptualize where description may fit into work practices. I examine subject cataloging and archival practices alongside social tagging as a means of drawing conclusions for possible new paths in integration.

1. Introduction and Background
While digitization has garnered much attention, the process is only a first step in shaping new forms of access. Last year, Wired writer Kevin Kelly forecasted that within the next decade, “each word in each book [will be] cross-linked, clustered, cited, extracted, indexed, analyzed, annotated, remixed, reassembled and woven deeper into the culture than ever before... every bit informs another; every page reads all the other pages.” This enthusiasm for the possible future has been accompanied by intense scrutiny of and disenchantment with traditional indexing and descriptive practices. The extent to which bibliographic description has been subject to scrutiny by library practitioners (Calhoun, 2005) and outside observers (Weinberger, 2007) calls into question traditions of cataloging and indexing, proposing “crowdsourcing” over control. The library community at large is beginning to re-examine cataloging practice and integration of bibliographic structures into the “fabric” of the Web.

In this paper, I will attempt to redirect this scrutiny, and apply some of the gains of this activity to the act of archival description and to the reality of description in Libraries, Archives and Museums. In recent years, scholars such as Duff, Yakel, Tibbo, Cox and Piggot have examined the archival process, applying new directions in description, and articulating user desires for finding materials across institutions. Drawing on this work, I examine context and comparison on the new web. Efforts in metadata configuration and interoperability have a goal of cross-institutional finding solutions (Prom, 2003), but many questions remain as to what these solution may look like for users and how they may function for LAM professionals.

2. Archival Practice: Individual, Institutional and Contextual
Thedore Schellenberg wrote in 1965, “An archivist should not classify records... the material he deals with does not lend itself to classification”. For him, the hierarchy and physicality of library classification can be inappropriately applied to documentary material. Instead, he proposed that the “multiple relations” of archival record groups “be shown more easily and more usefully in descriptive documents”. Subjectivity is not presented as a problem for archivists, but a non-issue to be avoided by adherence to established order: “An archivist should not attempt to develop a classification system at all”, writes Schellenberg, “he should simply establish... units or groups among the material and arrange such units in the order of their establishment”. The “descriptive documents” Schellenberg prescribes are, with few variations, the contemporary finding aid, and the main elements of such standards as Describing Archives, a Content Standard (DACS), Rules...
of Archival Description (RAD), and General International Standard Archival Description (ISAD(G)).

The social and historiographic models for Schellenberg’s archives are established, mainly governmental, agencies, a view that fails to explain or consider the upheaval in the historical profession(s) and the writing of history that began in his era (Higham, 1989). Schellenberg’s descriptive paradigm is rigid, and to this researcher, his rhetoric is outdated and direction lacking nuance. However, a larger question looms as to the appropriateness of this descriptive form for collections that fall outside of paper records of the United States Government, which may have more complex structures and actors. When a rigid descriptive template is applied widely, often this application is, as Richard Cox terms, “creative” (1990). Even if the current model fits ecumenically, we can at least explore how contemporary archival context may differ from Schellenberg’s rationality.

Rigid, outmoded instructions ensure idiosyncratic and highly subjective execution of archival description. Ideally, these descriptions would be locally informed and thus useful in context. Toms and Duff’s (1997) studies of archives users, however, point towards a preference for consistency, especially in language application. Practitioners have argued against freeform descriptions, saying that Schellenberg’s model is unmanageable on a large scale (Greene, 2005).

To adherents, this system stages information as evidence; to detractors, it is unworkable and unsuitable for contemporary information practices. This dilemma is very similar to the perceived weaknesses of any individually-based descriptive system. (See tables) Without clear definitions, standard generation techniques, or consistency of intent, access to archival materials is dependent on inferred context, intermediation of reference archivists, and luck; this bodes ill for both materials in original physical format and digital reproductions (Tibbo, 2003).

3. Bibliography, Social Tagging, and Assumptions

Schellenberg’s assertion that bibliographic measures were inappropriate for archival records is not without justification, yet it is also important to contextualize modern bibliography on an appropriately broad scale. Global cataloging standards were developed—and still rest—on an assumption of books as editioned, single author works used in print-volume context; there is little room in this scheme for uniqueness, physicality or perceived differences in representation. It is evident in bibliographic practice that Cutter’s “cult of the title page” persists, but does this make for the best description of digitized materials? The catalog record for a book as a whole rarely reflects the particular content of a single page, just as a finding aid does not represent a single document in an archival collection.

Despite much effort put into the linking of data, the majority of metadata standards are designed for single-item descriptions, and offer no real solution for linking parts to wholes. On the web, archival materials are not marked as unique, monographs are not easily identified as editioned, and any judgment as to the physical or situational context is almost always inferred. The blurry differentiation of context is a frequent issue for digitization guidelines and will continue to be so into the near future.

1. The tradition of literary manuscript cataloging has a parallel history to Schellenberg’s model. However, for the purposes of this paper, I do not make the explicit distinction. Over the past decade, major literary collections in the US have been active in adopting EAD, thus bringing their finding aids closer to archival norms.

2. To this point, application of large-scale vocabulary structures for Encoded Archival Description standards has been consistently problematic, as EAD was designed as an extremely permissive standard based upon the inconsistencies in archival practice.
Kelly and Weinberger argue that “crowdsourcing” description is the most direct means of linking data and for maintaining order in digital documents at large. Social tagging, as Tennis states, is an “indexing of individual craft interaction” independent of outside authority. Much as archival descriptions are written from the view of the archivist (albeit masked by protocol), and subject headings are plucked from LCSH by an individual cataloger, social tagging employs a personal warrant. However, tagging requires no assumption of authority, and no placement within an institutional context. For Schellenberg and his government records, this is abhorrent. But for archival collections that represent a multiplicity of viewpoints and have been researched by a diverse community of scholars, this represents an unforeseen opportunity to reflect the potential of collections in description.
Table 2. Discourses of Description
Altered significantly from the original, this table outlines components of archival discourse and the relations implicit in indexing work in subject cataloging and social tagging. Provenance and authenticity, which figure heavily in archival discourse, in turn play a minor role in the discourse of subject cataloging and a greatly unexplored one in social tagging. In comparison, the established, unquestioned stance of catalogers draws out the intricacies of the archival method.

<table>
<thead>
<tr>
<th>Discourse</th>
<th>Archival Description</th>
<th>Subject Cataloging</th>
<th>Social Tagging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authority</td>
<td>Institutional</td>
<td>Institutional (local and national)</td>
<td>Personal</td>
</tr>
<tr>
<td>Creator role</td>
<td>Inherent</td>
<td>Occult</td>
<td>Confessed</td>
</tr>
<tr>
<td>Application of principles</td>
<td>Interpretive, determined by institution and adapted standards</td>
<td>Generally routinized, shaped by the institution</td>
<td>Generally un routinized, matter of sense-making</td>
</tr>
<tr>
<td>Links between materials (intertextuality)</td>
<td>Record grouping, subject grouping, by date of accession</td>
<td>LCSH, other books in catalogs, other titles, user logs, user reference interactions</td>
<td>Collection of tags, others’ tags, and other web pages in collection-explicit in interface</td>
</tr>
<tr>
<td>Scope</td>
<td>Local/professional discourse</td>
<td>Totalizing discourse</td>
<td>Local discourse: social/group</td>
</tr>
<tr>
<td>Voice</td>
<td>Authoritative, scholarly</td>
<td>Authoritative, neutral</td>
<td>Personal</td>
</tr>
<tr>
<td>Perceived interpretation</td>
<td>Objective and informed</td>
<td>Objective and informed</td>
<td>Subjective and accountable</td>
</tr>
<tr>
<td>Projected audience</td>
<td>Researchers</td>
<td>Users</td>
<td>Self and others like self</td>
</tr>
<tr>
<td>Manifestation of Provenance</td>
<td>Administrative note, collection title</td>
<td>“A luxury”</td>
<td>Implicit</td>
</tr>
</tbody>
</table>

Table 3. Framework Components
In response to the changes made to the previous table, this table presents the archival discourse on the collection record alongside the parts and purpose of social tagging and subject cataloging.

<table>
<thead>
<tr>
<th>Components</th>
<th>Archival Description</th>
<th>Subject Cataloging</th>
<th>Social Tagging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purpose</td>
<td>Intellectual control</td>
<td>Fulfill Cutter’s objective #2</td>
<td>Share, innovate organization</td>
</tr>
<tr>
<td>Predication</td>
<td>Finding aid fields, metadata fields</td>
<td>Subject headings lists in an OPAC</td>
<td>Tags, Profiles, folksonomy connections</td>
</tr>
<tr>
<td>Function</td>
<td>Establish authenticity; serve as institutional record</td>
<td>Find and Collocate (formal and intentional)</td>
<td>Share (social or accidental)</td>
</tr>
<tr>
<td>Context</td>
<td>Archive, collections, researchers</td>
<td>Library, collections, users</td>
<td>The web</td>
</tr>
<tr>
<td>Format</td>
<td>Paper, reproduced on web</td>
<td>OPAC</td>
<td>Web document</td>
</tr>
</tbody>
</table>

4. Frameworks and Findings
Derived from Tennis’s 2006 rubric for comparison of social tagging and indexing and further informed by the archival literature mentioned above, these frameworks analyze the descriptive work processes, discourses, framework components, and structures of description for the three processes. In altering the tables, I compare the different practices on the basis of their field use, with emphasis on potentially divergent applications.

In comparison, none of these discourses emerges as representative of all types of searching or methods of research. Each appears particular, contextualized within a parameter of defined tasks or context. Moreover, their genesis in social tagging highlights the extent to which cataloging or archival work serves a professional perpetuation practice, dependent on intermediaries familiar with the discourse to translate user requests into search queries or browsing parameters.
Table 4. Structures of Description

For further comparative analysis, I expanded Tennis’ characteristics framework to represent the three practices in terms more amicable to archival process. Like in the original, this framework can serve as the basis of comparative statements on description and access.

<table>
<thead>
<tr>
<th>Structures of Description</th>
<th>Archival Description</th>
<th>Subject Cataloging</th>
<th>Social Tagging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Role of vocabulary</td>
<td>Uniformity</td>
<td>Comprehensiveness</td>
<td>Ease in recall</td>
</tr>
<tr>
<td>Vocabulary or terminology (warrant)</td>
<td>Institutional/ standards (RAD, DACS, EAD)</td>
<td>LOC’s warrant</td>
<td>Personal information warrant</td>
</tr>
<tr>
<td>Format specific?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Implication of provenance</td>
<td>Record of custodianship</td>
<td>Not under normal circumstances</td>
<td>Relational</td>
</tr>
<tr>
<td>Guidelines for narrative</td>
<td>Implicit, by example</td>
<td>LCSH, AACR2</td>
<td>None</td>
</tr>
<tr>
<td>Degree of Control</td>
<td>Institutional, personal</td>
<td>Institutional</td>
<td>Personal</td>
</tr>
<tr>
<td>Uniformity of fields</td>
<td>Not consistent</td>
<td>Consistent</td>
<td>Not consistent</td>
</tr>
<tr>
<td>Levels of Hierarchy</td>
<td>Varies with arrangement</td>
<td>Varies with pre-coordination</td>
<td>Not present</td>
</tr>
<tr>
<td>Purpose(s)</td>
<td>Authentication, intellectual (vs. physical) control</td>
<td>Collocation and precision</td>
<td>Management and sharing</td>
</tr>
<tr>
<td>Perceived accuracy</td>
<td>Accurate, objective</td>
<td>Accurate, objective</td>
<td>Associative</td>
</tr>
<tr>
<td>Syndetic Structure</td>
<td>Partial</td>
<td>Partial</td>
<td>None</td>
</tr>
<tr>
<td>Type of Control (policy?)</td>
<td>Local policy and descriptive standard</td>
<td>Local Policy and LCSH policy</td>
<td>No policy</td>
</tr>
<tr>
<td>Coordination of records</td>
<td>Mainly postcoordination</td>
<td>Precoordination, post-coordination</td>
<td>Varies, on individual level</td>
</tr>
</tbody>
</table>

Tennis reminds us that social tagging is incipient, and that indexing has yet to be updated for contemporary technology. Finding aid interfaces, much like indexes, have yet to be reinvented for web use. Instead, they are formatted and transcribed for identical translation to their paper forms, based for the most part on Schellenberg’s components. Like indexing, archival description languishes because of its fixed nature. Without a conceptual mapping of parts, aspects, perceptions, and representations of material, information systems will be unhelpful and eventually obsolete. Instead of adding value to their materials, they will detract from them.

5. Conclusion

In order to propose a model of descriptive practice that will serve users across institutions, one future direction might be to fully investigate these tools and their uses as information environments. As stated before, contextual clues still very much indicate that finding aids, and to a lesser extent catalog records, imply an intermediary. As physical presence is no longer possible at all times, objects’ digital homes must have the same indicators of direction that their physical ones continue to do.

For those who design and maintain descriptive systems, it is imperative to examine not only the dimensions of systems, but the behavior facilitated by system parameters. If the public is as eager to describe materials as some allege, examining implicit restrictions on usage behavior is a critical step in cultivating user contributions. The recent flurry of user contributions to the Library of Congress’ Flickr photos suggests that much activity can ensue in a space that encourages it.
It is of interest to all information institutions to fully consider the materiality of their digital presentations and the translation of their institutional intent to the web. The work of MacNeil (2006) explores dimensions of authenticity in the digital environment. But a fuller investigation of other presentation aspects, such as collocation and authorship, is necessary to realize the implications of digital use. Material description and user context of digitized archives may be similar to that of their bibliographic counterparts, both for lack of physical barrier and presentation. Now is the time to capitalize on interest and outside investment in a digital future for cultural materials and their users.

References


Piggot, M. (2007). Human behaviour and the making of records and archives. Archives and Social Studies 01, no. 0


Jenn Riley — Indiana University

Moving from a Locally-developed Data Model to a Standard Conceptual Model

Abstract
While the LIS community has recently begun to explore conceptual models on a larger scale than in the past, the connection between these conceptual models and system functionality based on a given metadata element set is still largely unclear. This paper summarizes some recent developments in conceptual models in the LIS sphere, considers the effects on interoperability between systems for element sets with explicit rather than implicit conceptual models, and describes work by the Variations team at Indiana University to apply the lessons learned from conceptual modeling efforts to improve the interoperability of Variations metadata.

Conceptual models and metadata element sets
The Library and Information Science (LIS) literature contains a significant body of work describing the development and usage of descriptive metadata element sets such as MARC, MODS (Metadata Object Description Schema), and Dublin Core. There is also a noteworthy body of literature on conceptual models, which, broadly defined, includes such models as the Functional Requirements for Bibliographic Records (FRBR) for bibliographic information (IFLA Study Group 1998) and the CICOC Conceptual Reference Model (CRM) for cultural heritage documentation (International Council of Museums 2006). Despite the considerable work done in these two areas, little has been written about how they interact—to what degree various metadata element sets are implementations of standard conceptual models and what the effects on interoperability of conforming or not conforming to a conceptual model might be.

In fact, the very boundaries between abstract models, conceptual models, metadata element sets, and other relevant standards such as content standards are still not well understood. Those documents describing themselves as conceptual models operating in the digital library metadata sphere are structured in a variety of ways. FRBR is a human readable description describing the results of entity-relationship analysis employing conceptual modeling but not full conceptual analysis (Renear et al. 2006), whereas CIDOC CRM is a formal ontology. The Dublin Core Metadata Initiative (DCMI) Abstract Model (Dublin Core Metadata Initiative 2007) is not a conceptual model like FRBR or the CIDOC CRM, but rather operates at a higher level of abstraction. It describes itself as an “information model,” and uses UML diagrams to illustrate the formal model, supplemented by textual descriptions. FRBR and the CIDOC CRM in contrast define specific entities and their relationships rather than generic constructs intended to be used with any type of data. Each of these models performs a common function, however, in defining a way of thinking about data, and abstract and conceptual models can be used to some (possibly limited) degree interchangeably when thinking about metadata interoperability.

Metadata element sets suffer from a similar lack of clarity of definition. While these generally provide what most think of as “fields” for information, some also include prescribed or recommended methods for constructing the values for those fields, a function often discussed as the realm of content standards or encoding schemes (which includes controlled vocabularies). The distinction between these concepts is useful to make, but it is just as important to realize that as a practical consideration, a given standard may cross the boundaries between them in places.
By their very nature, the various “buckets” for data defined by a descriptive metadata element set represent a certain way of thinking about resources. That is, “underlying most metadata specifications there is an assumption about an abstract model…[which] specifies the concepts used in the standard, the nature of terms and how they combine to form a metadata description” (Nilsson et al. 2008). Both abstract and conceptual models can be either explicit or implicit. Most element sets in the library community fall into the implicit category—there is no formal, explicitly defined conceptual model underlying MARC or MODS, for example. Metadata element sets in this implicit category are generally defined by their encodings, rather than formally specifying both a conceptual model and one or more bindings for that model into an encoding syntax. This general approach as taken in the library realm is illustrated by a definition presented in the American Library Association’s Committee on Cataloging: Description and Access Task Force on Metadata’s final report:

A metadata schema provides a formal structure designed to identify the knowledge structure of a given discipline and to link that structure to the information of the discipline through the creation of an information system that will assist the identification, discovery and use of information within that discipline. (American Library Association, 2000)

In this definition, the metadata schema is both the conceptual model (“knowledge structure of a given discipline”) and the encoding of that model into a metadata element set (“information system”). Greenberg’s definition of a metadata schema is reminiscent of the ALA definition, as a “conceptualization that is represented or formalized in a specification” (Greenberg 2003, 1878). Although Greenberg rightly states that a metadata schema can have multiple specifications, only recently has the idea of defining a conceptualization separately from its specification gained real traction in the library metadata community. The FRBR model, first formally published in 1998, is a good example, reflecting revolutionary thinking at the time of its writing that only began to be seriously considered for practical applications nearly ten years later.

The history of the Dublin Core Metadata Element Set (DCMES) illustrates the progression in thinking in this area, from an element set representing an implicit conceptual model to recognizing the benefits to interoperability and sustainability of specifying a separate, fully abstract, model. The DCMES was initially developed in 1995 (Weibel et al. 1995), expanded slightly in 1996 (Weibel and Miller 1996), then further expanded to include the notion of qualifiers in 2000 (Weibel and Koch 2000). It was only later that the need for a formal abstract model became clear, and the DCMI Abstract Model, heavily influenced by the RDF model, was first released in 2005. It “incorporates the concepts of metadata vocabularies, schemas, formats and application profiles into a single framework that can be used to analyze and compare metadata standards, and aid in the process of harmonization of metadata standards.” (Nilsson et al. 2006, 1) Once the Abstract Model was finalized, the DCMI was able to re-think its metadata encodings with the model in mind. New encoding guidelines for DC in RDF have been approved, and guidelines for XML and HTML/XHTML meta and link elements are under review and are expected to be approved some time in 2008 (Dublin Core Metadata Initiative 2008).

**Functions of a conceptual model in an information system**

Conceptual models, explicit or implicit, have a profound effect on systems that implement metadata element sets. Both system functionality and interoperability of metadata are governed to a large degree by the conceptual model underlying metadata sets used.
The effects of a conceptual model on system functionality are perhaps most easily seen in discovery capabilities. At the most basic level, data not present in the conceptual model underlying an element set is not available in search or browse for end-users. If a conceptual model does not recognize the notion of a genre for a resource, for example, this concept will not be available in element sets and encodings based on it, and therefore not available to end-users to either search on or view in search results.

Discovery functionality is also affected at more subtle levels. In the digital world, resources frequently exist in multiple versions. The degree to which a conceptual model makes clear distinctions and relationships between these versions affects which version(s) are described in a record created according to a given metadata element set. The distinctions made between versions in the FRBR Group 1 entities, for example, if carried through into an element set, would allow a system to group multiple versions of a resource together in some cases, and make distinctions between them in others.

Metadata interoperability is an ongoing challenge in information systems. In our diverse information environment, there exists a very real need for many different element sets designed for different types of resources, different types of institutions, and different approaches to description to co-exist. Transformations between these element sets can be effective when they are performed between those that conform to the same underlying conceptual model. For example, MARC to MODS transformations, while not lossless, result in reasonable-quality MODS records, because they are based on similar implicit conceptual models. Conversely, a generic mapping between MODS and VRA Core (Visual Resources Association 2007) is much less likely to be fully successful, as the work/image distinction central to the VRA Core is not present in the conceptual model underlying MODS. Similarly, it would be difficult to automatically map a MARC record to an element set based on the FRBR conceptual model due to the fact that, while many MARC elements are clearly related to only one FRBR entity, some could under different circumstances contain data that apply to different entities, without a machine-readable way to know which is being described.

Some research goes so far as to say “the notion of reusing ‘elements’ between metadata standards and formats using incompatible abstract models is fundamentally flawed” (Nilsson et al. 2007). In response to these concerns, the DCMI Abstract Model is an attempt to create a generic abstract model that could underlie any metadata element set, with the goal of increasing interoperability and making transformations between element sets and encodings conforming to the model more effective. Even the understanding of the same element set by two systems is potentially improved by the existence of an explicit abstract model: “when two applications want to exchange Dublin Core metadata, they understand metadata through the lens of the abstract model” (Nilsson et al. 2006, 3). One can imagine in this environment a future in which the choice of specific element sets and encodings would become much less important than it is today; conformance to the conceptual or abstract model would become the focus of planning, and the details of encoding syntax of lesser concern.

Applying these concepts in practice
The Indiana University Digital Library Program has been watching the developments in the area of conceptual models with interest, and took a lesson from the DCMI with regards to the benefits of revisiting encoding syntaxes with an explicit model in mind. We believe that adherence to an explicit conceptual model can provide many of the benefits the DCMI
is demonstrating can be gained through the use of a shared abstract model. The Variations3 Digital Music Library project at Indiana University represents a valuable illustration of these principles in a production-level system. Variations3 provides students and faculty at Indiana University with online access to streaming audio, scanned score images, and encoded scores from the University’s library collections, and is also being tested at several other institutions. In 2001, as part of a previous incarnation of this project (known as Variations2), the Variations team developed a data model for descriptive and structural metadata for discovery and use of objects in the system. The model was formally expressed as a data dictionary (Indiana University 2003), rather than as a conceptual model or encoding syntax. The data dictionary was then transformed by project programmers to an XML syntax, but at the time of original development no XML Schema or DTD was developed to govern that syntax.

![Figure 1. High-level view of Variations2/3 data model](image)

The lack of an explicit conceptual model and the definition of an encoding syntax in application code alone made it difficult to implement changes to the Variations data model over time. The drawbacks of our initial approach became clearer as the project moved into the Variations3 phase and began to focus on sustainability plans for the software and its metadata model. The Variations team began to question the long-term viability of a locally-built and informally specified model, and with lessons from DCMI in mind initiated an investigation of the desirability and practicality of converting our local data model to one based on FRBR. While the Variations data model bears a striking resemblance to FRBR, there are a few crucial differences.

![Figure 2. Variations 2/3 entities compared to FRBR Group 1 entities](image)

<table>
<thead>
<tr>
<th>Variations2/3 Entity</th>
<th>FRBR Group 1 Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work (more concrete than FRBR Work)</td>
<td>Work</td>
</tr>
<tr>
<td>Instantiation (can only appear on one Container)</td>
<td>Expression</td>
</tr>
<tr>
<td>Container (includes some copy-specific data)</td>
<td>Manifestation</td>
</tr>
<tr>
<td>Media Object (defined as a digital file)</td>
<td>Item</td>
</tr>
</tbody>
</table>
The FRBR model is gaining increasing traction in the library community, for example in modified form in OCLC’s Open WorldCat (OCLC), and underlying Resource Description and Access (RDA) (Joint Steering Committee 2008), the planned successor to the AACR2 content standard. The recent high-profile report of the Library of Congress Working Group for the Future of Bibliographic Control highlighted FRBR as a promising and potentially revolutionary development, while simultaneously issuing a strong call for its practical testing (Library of Congress Working Group, 2008). Given clear movement towards FRBR as the conceptual model underlying bibliographic data in the future, the Variations team developed a report (Riley et al. 2007) defining how FRBR concepts would apply to musical materials in our system. This report refines the definitions of the FRBR Group 1 entities as they apply to musical materials, provides guidelines on the distinction between the Work and Expression entities, selects attributes and relationships from FRBR that apply to musical materials, and defines a small number of new attributes and relationships not explicitly defined by FRBR but deemed necessary to adequately describe music. The development of this report convinced the Variations team that moving to a FRBR implementation was both possible and desirable, and would represent a major step towards ensuring interoperability of Variations data with that of the rest of the library community into the future. We were not alone in this conclusion. Gradmann, for example, has proposed that FRBR could function as “a kind of specific meta-ontology in the field of librarian information objects” (Gradmann 2005) — that is, as a conceptual model that could be instantiated in a specific encoding. The Variations team’s report operated on a conceptual level, and did not go as far as defining a data model or encoding for FRBR-based data; only after we were convinced that FRBR was the appropriate conceptual model for our system to instantiate were we ready to re-think the metadata element set used in the Variations system.

The encoding of the FRBR model into a metadata element set is by no means a trivial task, however—IFLA has not released a formal data model, element set, or encoding syntax for FRBR data, nor has any production-level FRBR implementer publicly released a data model or encoding used at any reasonable scale. Despite recent activity cementing FRBR concepts in the library community, potential implementers still do not have a shared, widely-supported, FRBR-based data model that can be implemented in the same way in multiple institutions.

A few options do exist, however. The Variations team will investigate each of these options in depth in summer 2008, and determine if any meets our functional and sustainability needs. The first is the “Expression of Core FRBR Concepts in RDF,” which grew out of the Semantic Web advocacy efforts of a software company with ties to the library community (Davis 2005). While an encoding syntax that works within the wider Web environment and not just within the library community is attractive, it is unclear to the Variations team at this time if an RDF approach is the right one, given its limited usage in libraries. A second option is the FBRRoo ontology, a product of work to harmonize FRBR with the CIDOC CRM (International Working Group 2008). The Music Ontology, a third choice, has recently arisen from work by Music Information Retrieval researchers. It is defined as an RDF OWL ontology, and builds upon existing frameworks such as FOAF and FRBR (Raimond et al. 2007). At first glance the wider coverage of both FBRRoo and the Music Ontology beyond the FRBR model appear to be out of scope for Variations, but more investigation is warranted.
It is possible that none of these existing encodings will provide a good match for the needs of the Variations3 project. In this case, our project team will embark on our own definition of a formal data dictionary and encoding syntax, perhaps defined as W3C XML Schemas, representing a data model explicitly tied to the FRBR conceptual model, and make the results of our work widely available for testing and implementation by others. Given other demands on development resources, we do not believe we will be able to perform a full-scale re-architecting of our system to a FRBR-compliant model as part of our current grant-funded project. Our approach will therefore be to complete planning work necessary for defining a FRBR-based encoding during our current project, and make any small-scale system changes necessary to implement this switch at a later date. We believe converting to a FRBR-based model is essential to the long-term sustainability of the Variations software. Given the increasing importance of formal conceptual models in the metadata landscape, the Variations team believes that our work will continue to make significant contributions to the state of the art by demonstrating in a production system the utility of metadata element sets that explicitly represent the FRBR conceptual model.

Acknowledgment
This paper is based in part upon work supported by the Institute of Museum and Library Services. Any views, findings, conclusions, or recommendations expressed in this paper do not necessarily represent those of the Institute of Museum and Library Services.

References


Jan Pisanski — National and University Library, Ljubljana, Slovenia  
Maja Žumer — Faculty of Arts, Ljubljana, Slovenia

How Do Non-librarians See the Bibliographic Universe?

Abstract

Functional Requirements for Bibliographic Records (FRBR) is a relatively new conceptual model of the bibliographic universe. While it is recognized among library experts, there is a considerable lack of user studies. A pilot study, consisting of three different tasks, was conducted to test the instruments for acquiring mental models of the bibliographic universe. Results show that users do not have a consistent mental model of the bibliographic universe and that various techniques used can be useful for acquiring individuals’ mental models of the bibliographic universe. Of the three tasks, the one asking people to rank pairs of similar item according to substitutability revealed results that were closest to FRBR, while card sorting and concept mapping exercises failed to provide a single alternative model.

Introduction

Functional Requirements for Bibliographic Records (FRBR) is a conceptual model which represents a general view of the bibliographic universe (Functional… 1998). The model consists of bibliographic entities, relationships between them and attributes of these entities. Entities are divided into three groups. Group 1 consists of entities that are products of intellectual or artistic endeavour; Group 2 consists of entities that are responsible for intellectual or artistic content; while Group 3 is made up of entities that serve as subjects of intellectual or artistic endeavour. Group 1 entities are: Work, Expression, Manifestation and Item. Work is a distinct intellectual or artistic creation, expression is an intellectual or artistic realization of a work, manifestation is a physical embodiment of an expression, while item is a single exemplar of a manifestation.

While FRBR caused some discussion in library circles, the implications of its application are still largely unknown. Even the FRBR Final Report recognizes that the model is not meant as perfect or unchangeable, and theoreticians and practitioners alike (Delsey 2005; Kilner 2005; Hickey, O’Neill 2005) have expressed ideas how to improve on the current model.

While a handful of FRBR-based, or rather FRBR-like, implementations exist and early studies show enthusiasm for frbrized catalogs on part of the end-users (Kilner 2005; Jepsen 2007) we cannot really claim to know much about how users view the bibliographic universe. Although there had been many user studies prior to production of FRBR, no user studies were conducted during the course of FRBR creation. In fact, the highly controversial proposal to suspend work on RDA (Resource Description and Access) made by The Library of Congress Working Group on the Future of Bibliographic Control was motivated by the same recognition. While it may not be the most diplomatic thing to suggest, it highlights the importance of the issue.

We propose to rectify this apparent void by eliciting individuals’ mental models of the bibliographic universe. While this may seem redundant as experts certainly are the ones who know their area of expertise best, there are good reasons for conducting such a study. Firstly, there is no common agreement on the validity of the conceptual model among the experts. Secondly, users are the ones who are supposed to benefit from FRBR, but only as far as the conceptual model matches their mental models. Of course, users may have imperfect mental models or have more than one mental model of the bibliographic universe.
and these mental models may not be stable. But that is exactly why the library community should look into mental models of their users.

To this end, we conducted a pilot study. The purpose of our study was twofold: to test the applicability of this approach, as well as to possibly practically point out any major theoretical flaws in FRBR. The methodologically closest research to ours was work done or proposed by Carlyle (1999; 2001; 2003), relating to sorting of various manifestations with the same progenitor and substitutability of material. However, in her card-sorting study Carlyle was interested in various sorting criteria on the same level, whereas we were interested in the single criterion that arranges bibliographic entities into different levels, and her proposed research focused on real-life library user needs, while we deliberately tried to avoid any reference to catalogs or even libraries as we were interested in people’s mental models of the bibliographic universe. However, it would be wrong to assume that people do not form their mental models of bibliographic universe at least partially based on their experience with library catalogs.

According to Norman (1998), mental models are models people have of themselves, others, the environment and the things they interact with. Mental models are formed through experience, training and instruction. They are an internal representation of the outside world. Mental models vary with time and the knowledge and experience acquired. Usually, there is a tendency for user’s mental model to become closer to conceptual model. Research usually focuses on conditions where expert’s conceptual model of a system is in place, while user’s understanding of the system is to be determined. While it is true that in the case of bibliographic universe the FRBR conceptual model is in place, we would like to check the validity of the model which is normally a given. Mental model research is not uncommon in LIS, e.g. Michell and Dewdney (1998) identified 42 such projects.

The Design of the Study
The study was conducted on 15 participants, residing in or in vicinity of Slovenia’s capital, Ljubljana. Participants of our study were asked to complete three tasks: card sorting, making concept maps and comparing/contrasting pairs of books. These three tasks were chosen to get a more well-rounded view of mental modes of the bibliographic universe.

Specifically, the first task asked of the participants to sort cards containing various instances of FRBR group 1 entities into at least three groups based on the criterion of concrete/abstract (physical/non-physical) nature. We asked for at least 3 groups to avoid having respondents splitting cards into just general “abstract” and “concrete” groups. Card sorting was performed on two separate examples (Hlapec Jernej in njegova pravica and The Da Vinci Code), in order to detect any inconsistencies. Also, participants were asked to name/describe groups in an attempt to find more user-friendly names for the entities. Participants were instructed not to pay attention to the descriptions but rather to what those descriptions were of or represented. As this is a rather subtle distinction, most of the participants had trouble understanding the difference. There were 13 cards used for the example of Hlapec Jernej in njegova pravica and 14 for The Da Vinci Code, which also expanded into the territory of motion pictures and audio books. Descriptions were worded in such a way that they would not remind participants of catalogs and we also tried to avoid using exactly the same attributes, as to not get results based on similarity of descriptions. Therefore cards would be as diverse as “Ivan Cankar’s Hlapec Jernej in njegova pravica”, “The copy of the book The bailiff Yerney and his rights from 1930 held in the British Library”, “Croatian translation by Stanko Tomašić called Sluga Jernej i nje-
govo pravo, which has 94 pages” and “Cankar’s original text of Hlapec Jernej in njegova pravica”. The last description was found to be most problematic to understand, as some deemed it to be somehow only connected to the dramatization, which was also part of the example, some thought it denoted a draft, some thought it referred to Cankar’s manuscript and some could not place it at all, as we saw in Task 2.

For the second task, participants were given the same cards used in card sorting for Hlapec Jernej in njegova pravica and asked how these cards are inter-related and, more specifically, “What comes out of what?”. They were told not to necessarily think in terms of chronological order and that the resulting graphs need not be linear. Again, participants were asked to describe the resulting “concept maps”. Almost all of the participants changed their mental models during the course of description, sometimes as a result of researcher asking questions to make things clearer, but it was mostly fine tuning.

The third task consisted of two parts. First, the participants were asked to describe 11 pairs of similar items (mostly pairs of books, but in two cases also a book and a DVD) presented to them, state how they are similar and whether they are substitutable. It was explained to the participants that substitutable was meant in terms of: “If one cannot acquire one item in the pair, would one be satisfied with the other?”. It is important to note that participants were instructed to name any possible condition for the two items not being substitutes, regardless of whether or not it applied in the case of particular participant. Finally, they ranked the pairs according to the degree of similarity / degree of substitutability. Although the task allowed for the two different rankings none of the participants did so.

Each card sort, conceptual map and ranking was photographed with a camera and all of the explanations were recorded on a sound recording device.

Results

Our research showed that, in general, the participants did not have a ready-made mental model of bibliographic universe. In fact, not only did mental models seem to change with the task, but also using different examples within the Task 1. It must be said that this may partly be attributed to the relative complexity of the tasks.

Most of the participants found the first two tasks difficult, asked for further explanation and often expressed their dissatisfaction with the criterion or the design of the tasks. In fact, the naming of the criterion for Task 1 was found to be critical during the design stage of the study and much consideration was given to the appropriate naming of the criterion. If participants asked for more explanation, they were told that a book can be a concrete physical object, but it can also be something more abstract.

All but two participants failed to sort cards according to the criterion given and rather sorted them using their own criteria. In these cases the criteria used most commonly were language and physical format, as well as a combination of both with the original criterion. In Task 2 some participants only made vague connections between larger groups of cards. It is also safe to assume that some participants did not read the cards carefully enough, although they were initially instructed to do so. In fact, that was the reason given by individual participants, when apparent inconsistencies were detected. In Task 3, some participants failed to consistently comment aloud, again contrary to the instructions and despite repeated prompting. Seemingly, participants failed to note the obvious (for instance, titles of books were seldom stated by the participants), as they probably found it trivial. Also, although there was no time constraint, some participants seemed to hurry through the tasks.
None of the individual card sorts and conceptual maps obtained in the first two tasks was completely in accordance with FRBR, although 10 of the 15 concept maps captured the Work—Expression—Manifestation progression or the Expression-Manifestation-Item progression at least once, and 5 even captured the whole Work-Expression-Manifestation-Item chain, but not consistently (one such example is found in Figure 1, where participant U1 established Work-Expression-Manifestation-Item progression from work through original Slovenian text and first edition to copy of that edition). In fact, no clear model emerged from the first two tasks, as mental models elicited shared relatively few common elements. However, in Task 2 six of the participants, including U1 in Figure 1, considered the original text of a work relatively important compared to other FRBR expressions (this might have been influenced by the example of Hlapec Jernej in njegova pravica by Ivan Cankar, considered a classical work in Slovene literature). Four others even put all entities pertaining to the original at the top of their hierarchies. Also, a dramatization of a work, which is considered a separate work in FRBR, was generally not considered as such by the participants, when eliciting their mental models. However, dramatization was indeed branded as a separate work by two participants in informal conversation. An example of a concept map can be seen in Figure 1 (the descriptions here are used for clarity and are not those actually used on the cards).

In Task 3, all of the participants stated that when two items were in different languages, those who do not speak both languages would not consider them as substitutes. That was the only element of non-substitutability that vast majority of participants explicitly agreed on. However, for example, most participants did agree that a book and a movie or two versions of a textbook published twenty years apart, generally speaking, are not substitutes but they named various reasons (e.g. content, carrier or both for book/movie; age, edition, content, or some combination of these for textbooks), which were in some cases difficult to distinguish. In light of apparent importance of original text detected in
the first two tasks, it must be stressed that in Task 3 most participants indicated preference for newer texts when comparing two versions of the same textbook.

Also, while there was some variation, a general pattern emerged from the ranking exercise, as seen in Table 1. Two practically identical items were found to be most substitutable (Charter of Parma, Uncle Tom’s Cabin). Next, there were different manifestations of the same expression (Okus po grenkem, Master and Maragarita, The Mystery of the Blue Train), followed by different expressions (Where to in Ljubljana?, Skrvnost modrega vlaka, Out of Africa) and pairs containing a book and a movie, based on that book, (Poirot, A Room with a View) were found to be most dissimilar (although participants rarely indicated whether the distinction was due to differences in content, carrier or both). This generally follows the classic FRBR Item-Manifestation-Expression-Work structure.

It is interesting to note that 6 of 11 pairs were considered to be the most similar by at least one participant and to be the least similar by at least one other participant, although it must be said that one participant only made two separate groupings. The one pair that varied the most in rankings was a pair of textbooks (Economics) separated by more than 20 years. This can be attributed to the cultural differences mentioned in the FRBR report, although, as far as users are concerned, apparently these differences exist even between neighbours or members of the same family.

Table 1. Average ranks of pairs according to degree of substitutability

<table>
<thead>
<tr>
<th>Work name</th>
<th>Most notable differences</th>
<th>Avg. rank</th>
<th>Var.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Charter of Parma</td>
<td></td>
<td>2.233</td>
<td>1.067</td>
</tr>
<tr>
<td>Uncle Tom’s Cabin</td>
<td>Owner’s signature</td>
<td>2.3</td>
<td>0.957</td>
</tr>
<tr>
<td>Okus po grenkem</td>
<td>Format</td>
<td>3.4</td>
<td>0.828</td>
</tr>
<tr>
<td>Master and Maragarita</td>
<td>Number of volumes, Format, Age</td>
<td>3.867</td>
<td>1.088</td>
</tr>
<tr>
<td>The Mystery of the Blue Train</td>
<td>Additional photos and interviews, Age, Images from movie</td>
<td>5.333</td>
<td>5.702</td>
</tr>
<tr>
<td>Where to in Ljubljana?</td>
<td>Language, Slightly different contents</td>
<td>7.267</td>
<td>1.817</td>
</tr>
<tr>
<td>Skrvnost modrega vlaka</td>
<td>Language, Age</td>
<td>7.33</td>
<td>1.845</td>
</tr>
<tr>
<td>Out of Africa</td>
<td>Language, Additional short story, Author’s pen name, Age, Image from movie</td>
<td>7.5</td>
<td>6.464</td>
</tr>
<tr>
<td>Economics</td>
<td>Contents, Age, Additional author, Owner’s signature</td>
<td>7.533</td>
<td>1.409</td>
</tr>
<tr>
<td>Poirot</td>
<td>Medium, Contents, Number of stories, Age</td>
<td>9.5</td>
<td>2.678</td>
</tr>
<tr>
<td>A Room with a View</td>
<td>Medium, Contents, »Demanding literature«</td>
<td>9.733</td>
<td>1.102</td>
</tr>
</tbody>
</table>

Therefore, while the results of the first two tasks were not particularly FRBR-supportive, participants did indeed rank non-borderline cases in the order that closely shadowed FRBR in Task 3, which was the most concrete of the three, as it not only provided the participants with real-life pairs of items, but also stimulated them to think in terms of material use. For the borderline case of textbooks (separate expressions of the same work or two separate works?), no clear answer was obtained; some even considered them as separate manifestations of an expression. However, on average the textbooks were placed together with other expressions, just like FRBR suggests.

Also two editions of Agatha Christie’s The Mystery of the Blue Train were on average ranked higher than other cases where there were differences in editions. Part of the reason was because one contained additional pictures and interviews with the actors from the movie of the same name. However, some, particularly those who could not speak English well, may have also considered them less substitutable because they were in English, not Slovenian like the other two examples.
Conclusion
Based on our research, we can say that this approach is both feasible and needed, although further research would benefit from more focus and more homogenous groups of participants. While none of the tasks were found to be completely irrelevant for the purpose of mental model elicitation, the difficulty, as well as the number, of tasks, suggests taking a one-task-at-a time approach. Although, based on our research, one would have to be careful when drawing conclusions from results of a single task. Having more examples for each task would help clarify the results. Also, in-depth interviews with participants can shed new light on the results, as we found when talking to participants off-the-record. On the other hand, although there was no time constraint, some participants appeared uncomfortable with the length of the exercises and the fact that they had to search for details.

The described study is only a part of larger study, which should provide an even better understanding of how individuals view the bibliographic universe.

References
Design and Evaluation of Multi-viewed Knowledge System for Administrative Electronic Document Organization

Abstract
This paper describes part of a current research investigating the feasibility of creating a faceted and multi-viewed knowledge organization system (KOS) for administrative document organization in online environments. Preliminary findings support the faceted and multi-viewed classification as a promising alternative to the hierarchical paradigm for personal administrative electronic documents organization. Further analysis about identification of semantic relations between facets to reduce number of facet descriptors is required. Technical improvements are also needed to enhance the faceted navigation interface used within the pilot test.

1. Introduction
Most operating systems support the organization of electronic records according to hierarchical structures or directory trees. It has not yet been proven that using hierarchical classification is appropriate for the organization of administrative electronic documents under employees’ custody (Dourish et al. 2000). This intuitive organization method can also be very constrained, since documents are only classified in one place according to one person’s logic, which is often “hermetic” for other employees (Jones & Maier 2003). Furthermore, the inherent boundaries of such a structure require making arbitrary decisions on which specific criteria the classification will be based on, for instance, the administrative activity, the document type or its subject, although a document can have several attributes and require be classified in several classes.

The use of a faceted and multi-viewed knowledge organization system (KOS) would eliminate some of these well-known constraints and overcome the fact that traditional hierarchical classification schemes draw on pre-coordinated representation languages. In this paper, the concept of multi-viewed KOS is used to designate the combination of formal faceted classification and user-based semiotic ontologies.

The faceted approach is based on the pioneering work of Ranganathan (1965), and has been extended in the work of Gödert (1991) and Ingwersen & Wormell (1992) among others researchers, and applied to information organization and retrieval in networked environment. As demonstrated in a previous work on the classification of administrative paper-based documents (Hudon & Mas 2001) and electronic records (Henderson 2005) this approach appears to be promising.

In parallel, there were many efforts from the knowledge engineering field to reflect “knowledge” of a domain (Zacklad et al. 2007) by using a multipoint perspective. Particularly, the “socio-semantic Web” approach advocates using semiotic ontology systems that characterize the documentary collection attributes with topics across several complementary or competitor point-of-views built by human actors.

Multi-viewed KOS which applies to the personal organization of electronic administrative documents has never been studied. As far as current research leads us, there is no comparable study in this area. This paper describes part of a current research funded by
the Fonds québécois pour la recherche sur la société et la culture (FQRSC)\(^1\), investigating the cognitive and technological feasibility of replacing traditional hierarchical classification schemes with a multi-viewed KOS for personal electronic document organization.

A three-phase methodology was used for the development and evaluation of a personal multi-viewed KOS: (1) *Personal Folder Names Analysis*: the analysis of folders names used by employees in a university setting to organize and retrieve electronic documents on their workstation; (2) *Facet Selection and Taxonomy Development*: the selection, naming and organization of documentary facets into taxonomies, and (3) *Multi-viewed KOS Evaluation*: the validation and refinement of the taxonomies within a pilot test.

2. Personal Folder Names Analysis

As a first step towards identification and selection of facets, we analyzed folder names created by employees for the organization of electronic documents under their custody. The objectives were to investigate the concepts used by participants to name folders and to identify commonalities in the types of folders created.

We used a sample of twenty-one personal classification schemes (i.e., folders structures) designed and used by French speaking administrative secretaries and internship coordinators at University of Montreal. These personal classification schemes were analyzed as part of a larger study which purpose was to assess if it was possible to find a document with the same effectiveness, regardless of the classification scheme used (Mas 2007).

Semi-structured interviews were carried out with administrative secretaries and internship coordinators. We asked participants not to tidy before the interview. Each participant was met individually at his workplace, in front of his main work computer. The purpose of these interviews was to gather, in the presence of the participant, quantitative data (e.g., total number of classes), qualitative data on the folder structure they designed (e.g., content description of their electronic directories) and finally, information on the type of activities performed by the participant.

Each folder structure was collapsed and transposed into Excel-based tables for analysis. Qualitative data relating to the criteria used for the logical division applied at the first two levels of each classification structure was obtained through manual examination and interpretation.

Through our analysis, we found out that the examined personal classification schemes presented various organizational approaches. An analysis of the concepts represented in the folders names sample revealed eight main criteria division used by participants to organize electronic documents: long-term *activity* (e.g., “Communications”), *topic* (e.g., “Plagiat”), *type of content* (e.g., “Base de données”), short-term activity or task like project (e.g., “Accueil des étudiants”), *time* (“sept. 2005”), *contact* (e.g., “Denis”), *institution* (e.g., “Faculté des Arts et Science”) and combination of *multiple* concepts (e.g., “Envoi Janvier 2004”).

According to the examined folder names there is no standard logical division criterion in the first two levels of the personal classification schemes. Consequently, we did not distinguish a significant pattern in the conceptual organization of the identified first two levels of folders structures. Meanwhile, *topic* is used as main principle of division

---

(28.78% of main folders) in ten classification schemes (47% of classification schemes), and reappears as a main concept at the second level (23.63% of sub-folders) in six of them (28% of classification schemes). In our sample, many folder names were made up of multiple concepts (19.23% of main folders, 18.40% of sub-folders), each representing two or multiple document dimensions (e.g., “Feuilles de temps Denis 2005”, “Dossiers étudiants 2004–2005”), this gives us an indication that the development of a faceted classification might be desirable.

The practice of mixing various concepts or principles of division in the development of a hierarchy is contrary to principles of classification because it creates classes of documents that are not mutually exclusives. Such a mix is found at the first two-levels in all examined personal classification schemes, making the structure less efficient for organization, retrieval and update (Molholt 1995).

3. Facets Selection and Taxonomy Development

Starting with the list of concepts identified through the folder names analysis, we extracted recurrent values (i.e., keywords and terms used in the folder names) (e.g., “Dossier étudiant”, “Faculté des arts et des sciences”) and clustered them into a limited number of classes already used by participants (e.g., Type of content) or into new classes revealed by the folders names descriptions (e.g., Recipient of document). Facets, sub-facets, and descriptors derived from these classes.

An analysis of the participants’ activities and tasks was also carried out. Based on process modelling and a review of the employee’s job description, major professional and administrative activities frequently or regularly performed by participant identified, decomposed and put into relation. Such a formal activity-based classification is known by records managers to simplify document organization and retrieval within their primary creation or context of use (Héon 1999).

Based on prior findings regarding Web faceted taxonomy usability, we restricted the number of top-level facets as well as the depth of the structure. As there is no documented best practice, we decided to restrict the number of first level facets to seven. The level of granularity has been limited to three levels with some exceptions going down to four levels.

During the taxonomy development, Ranganathan’s principles of facet analysis refined by Spiteri’s Simplified Model for Facet Analysis (Spiteri 1998) were used to determine the criteria for selecting and naming facets: differentiation (criteria of organization between top-level facets and their sub-facets are distinctive and logical), relevance (facets are expected to adequately reflect the purpose, subject, and scope of the classification system), ascertainability (facet names are simple and circumscribed), permanence (facets represent permanent qualities of the item being divided), homogeneity (Top-level facets are homogeneous, i.e., they’re situated at the same level of granularity), and mutual exclusivity (characteristics of division between facets are mutually exclusive, i.e. each facet describes one single aspect of a document).

Initially, before the first evaluation of the faceted taxonomy, the structure was composed of seven top-level facets (cf. Table 1): 1) Activité, 2) Type de document, 3) Session universitaire, 4) Provenance du document, 5) Destinataire du document, 6) Statut de l’étudiant, et 7) Liste des noms.
Table 1. Top-level facets

<table>
<thead>
<tr>
<th>Facet</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activité (Activity)</td>
<td>Refers to important or regular generic actions done on an annual basis. In this study, Activity was defined broadly to include long-term activities (e.g., Internship Management), short-term activities (e.g., project X) and events (e.g., Salon Studies).</td>
</tr>
<tr>
<td>Type de document (Type of content)</td>
<td>Refers to common properties about form and function of a set of documents (e.g., Report).</td>
</tr>
<tr>
<td>Session universitaire (Academic Session)</td>
<td>Refers to the academic year or session (e.g., 2003–2004, Fall2003).</td>
</tr>
<tr>
<td>Provenance du document (Origin of the document)</td>
<td>Refers to the name of entity (e.g., individuals, groups of individuals or institution) that creates or sends documents to the employee on a regular or occasional basis (e.g., Hospital X).</td>
</tr>
<tr>
<td>Destinataire du document (Recipient of the document)</td>
<td>Refers to the name of entity (e.g., individuals, groups of individuals or institution) that receives documents from the employee on a regular or occasional basis. Note: Descriptors of Origin of the document facet are reused here (e.g., Hospital X).</td>
</tr>
<tr>
<td>Statut de l’étudiant (Status of the student)</td>
<td>Refers to the status of the student with whom the employee exchange documents on a regular basis (e.g., “Inscrit”).</td>
</tr>
<tr>
<td>Liste des noms (Names List)</td>
<td>Refers to an authority list composed of the names of entities (e.g., individuals, groups of individuals or institution).</td>
</tr>
</tbody>
</table>

Thirty-eight second-level classes and 174 third-level classes were also available in the structure. The specific concepts Topic and Project were considered as classes of the generic Activity facet; however, the Contact concept was decomposed in two specific facets: Origin of the document and Recipient of document. The evaluation phase was aimed at determining whether the absence of Topic, Project and Contact facets from the top-level was disturbing for the user and allowing him to freely add these concepts, if required from his perspective.

4. Multi-viewed KOS Evaluation

If there is empirical evidence suggesting better performance by faceted queries (Vakkari et al. 2004), it remains an empirical question on how employees might interact with a faceted classification to organize their electronic records on a daily base. The purpose of this evaluation was twofold: first, to test the pertinence and validity of the selected facets; and second, to investigate the usability of a browsable and a collaborative tool for the faceted classification of electronic administrative documents in workplace.

A Web 2.0 oriented software solution was used to represent, validate and co-build the newly developed faceted taxonomy. Agorae is both an innovative approach and software platform developed by Tech-CICO team2. The Agorae “socio semantic Web” approach is usually used by a given community to design the representation of a domain cooperatively and find domain entities and collections by browsing “multi-points of views” (i.e., non-formal or heuristic attributes) (Cahier, Zaher & Zacklad 2007). Agorae proposes interactive features depending on three user roles: 1) Consult (user has a general view upon all facets, and can browse among classes and sub-classes), 2) Edit (user can create, delete, modify or move a facet or a class) and 3) Classify (user can describe a document by filling a text box, and by linking it to any class under one or more given facet).

The study adopted a naturalistic approach to investigate a user’s first interaction with the Agorae prototype to explore a faceted taxonomy to organize administrative document.

---

2. Equipe Technologie de la Coopération pour l’Innovation et le Changement Organisationnel (Tech-CICO), Université de technologie de Troyes, France.
One internship coordinator was invited to try out the Agorae interface for two weeks to organize any new electronic administrative documents created or received during the test period. The respondent was invited to make structural and conceptual changes from her own perspective with the Agorae software. At the end of the study, the newly developed multi-viewed KOS (made of formal faceted taxonomy and user’s new facets and classes) was examined and the respondent was interviewed in order to gain conceptual and technical appreciation.

Comments received on the conceptual design were generally very positive and encouraging. However, in complement of the formal facets proposed, we expected that new “non-formal” or subjective facets would emerge from the user-perspective but it wasn’t the case, probably because of the short test period. Meanwhile, the respondent made relevant comments and suggestions to improve the faceted taxonomy and interface. The following is a summary of comments and suggestions received.

The Activity facet was always used by the respondent for the organization of documents; the “Internship Management” class more often than others. From the respondent perspective, it seems that “all is Internship Management” but she didn’t want to delete this facet and its inherent classes considered as “Dossiers” for regrouping all documents relating to the same matter.

Since the Academic Session, Types of content, Origin of the document and Recipient of the document facets will eventually be linked to authority lists, the number of names available could be rather discouraging for the respondent who has to browse through a long list of descriptors. In order to better support the browsing process, a future interface should to narrow the list of types of content linked to a specific activity (e.g., “Curriculum Vitae” appears only with “Job Profile Management” and “Internship Management”) or a specific academic period associated to a type of document (e.g., “Student Guide” with “Fall”, which correspond at the beginning of the academic year). This semantic relation analysis between facets would improve effectiveness of the organization and search.

The number of facets used by the respondent for multiple-concepts document organization was also an interesting outcome of the pilot and highlighted the value of a faceted taxonomy for administrative documents organization. In the interview the respondent confirmed her need to organize and search documents through a combination of classes. During the evaluation period, the respondent assigned each document created or received to multiple conceptually simple classes to reflect all aspects of the document subject. Table 2 provides example of multiple allocations of a document to various facets and classes within and demonstrates the respondent’s understanding of the faceted classification.

According to the respondent’s feedback, the absence of Topic, Project and Contact as top-level facets didn’t seem to cause her difficulties. However, further research is now required to provide reliable and consistent data on this conceptual coverage and selection.

<table>
<thead>
<tr>
<th>Description</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curriculum vitae d’un étudiant non inscrit en stage envoyé à l’intention d’un milieu de stage à l’été 2003 (Hôpital X)</td>
<td>ACTIVITÉ: Gestion des stages TYPE DE DOCUMENT: Curriculum vitae SESSION ACADEMIQUE: Été 2003 PROVENANCE DU DOCUMENT: Étudiants → Étudiant X DESTINATAIRE DU DOCUMENT: Milieux de stages → Hôpitaux → Hôpital X STATUT DE L’ÉTUDIANT: Non-inscrit en stage</td>
</tr>
</tbody>
</table>
We also received technical recommendations on how to improve the Agorae software. Some of these comments were well known by the computer science community but remain technical challenges (e.g., reduce the number of steps to create or modify the name of a class). As already mentioned in previous knowledge organization research, it is difficult to get users to manually enter metadata. Employees are busy getting on with their work and are not usually concerned about managing their documents beyond the minimum required to ensure further retrieval. Not surprisingly, all ergonomic recommendations expressed by the respondent focused on the simplification and acceleration of the document description process common to most software solution.

4. Conclusion

Employees use personal classification schemes to facilitate access to their electronic documents. These personal classification schemes are likely to be hierarchical and very specific. Conceptual inconsistencies usually emerge, making the structure less efficient for organization and retrieval because of significant gaps in coverage.

Our research investigated the feasibility of creating a faceted and multi-viewed KOS for administrative document organization in electronic environments. The choice of the top-level facets for the taxonomy development was built on a compromise between a desirable low number of facets and an extensive conceptual coverage. Furthermore, the extensibility and structural hospitality of the newly developed faceted classification transposed in an innovative Web 2.0 browsing interface should allow to gradually expand the faceted taxonomy to reflect changes in the document collection to accommodate the user’s own perspectives.

Information overload, cognitive effort and work constraints are making document organization increasingly difficult. Technology should not only allow us to implement principles of faceted analysis and representation but better support the organization process by providing some level of automation when possible. The faceted structure is still evolving, and an analysis of the semantic relation between facets is now planned to improve the classification process. A longitudinal research design with more volunteers is also needed to validate ease of navigation, user friendliness, extensibility and flexibility of the faceted and multi-viewed KOS in an administrative work setting.

References


Abstract
Consistency studies have discussed the relationship between indexing consistency and exhaustivity, and it is commonly accepted that higher exhaustivity results in lower indexing consistency. However, this issue has been oversimplified, and previous studies contain significant misinterpretations. The aim of this study is to investigate the relationship between consistency and exhaustivity based on a large sample and to analyse the misinterpretations in earlier studies. A sample of 3,307 monographs, i.e. 6,614 records was drawn from two Chinese bibliographic catalogues. Indexing consistency was measured using two formulae which were popular in previous indexing consistency studies. A relatively high level of consistency was found (64.21% according to the first formula, 70.71% according to the second). Regarding the relationship between consistency and exhaustivity, it was found that when two indexers had identical exhaustivity, indexing consistency was substantially high. On the contrary, when they had different levels of exhaustivity, consistency was significantly low. It was inevitable with the use of the two formulae. Moreover, a detailed discussion was conducted to analyse the misinterpretations in previous studies.

Introduction
Ideally, if two indexers use the same thesaurus or classification system to index the same book, they are supposed to assign the same index terms or class numbers. In practice, indexers are not always consistent with each other, because subject indexing is essentially a subjective process. Since indexing consistency is an acceptable indicator of indexing quality, the increasing interest in enhancing information retrieval effectiveness has heightened the need for improving indexing consistency between indexers. Many factors can affect indexing consistency, among which indexing exhaustivity is the most important one and has been most extensively explored in former consistency studies. Considering the relationship between indexing exhaustivity and consistency, it is widely believed that higher exhaustivity leads to lower indexing consistency. However, the relationship between indexing consistency and exhaustivity is not as straightforward as it might seem. Actually, this issue has been oversimplified, and previous studies contain significant misinterpretations. The aim of this study is to investigate the relationship between consistency and exhaustivity based on a large sample and to analyse the misinterpretations in earlier studies.

Methodology
A sample of 3,307 monographs (i.e. 6,614 records) was selected from two Chinese bibliographical databases. One is the catalogue of the National Library of China (NLC), the other one is the catalogue of the China Academic Library & Information System (CALIS). NLC is the biggest public library system in China, while CALIS is the biggest academic library system. Both of the library systems utilize the Chinese Thesaurus (CT) as an indexing tool, which makes it convenient to comparing indexing results.

In earlier studies, the following two formulae were extensively used to calculate indexing consistency:

\[
CP1 = \frac{c}{(a+b-c)} \quad \text{(Hooper, 1965, reported by Leonard, 1977)}
\]
\[
CP2 = \frac{2c}{(a+b)} \quad \text{(Rolling, 1981)}
\]
where \( a \) denotes the number of terms assigned by one indexer, \( b \) denotes the number of terms assigned by a second indexer, \( c \) denotes the number of terms commonly assigned by the two indexers. In the interest of comparing with previous studies, both formulae are used for calculating indexing consistency in this study.

**Results**

The consistency of index terms between NLC and CALIS is calculated with the use of the two formulae described above. According to Hooper’s formula, the overall consistency of index terms is 64.21\%, while according to Rolling’s formula it is 70.71\%. A total of 18,182 index terms are assigned to the 6,614 records, i.e. 3,307 books, averaging 2.75 terms per record (Sd.1.20). The indexers of NLC totally assign 8,999 terms to these 3,307 books, averaging 2.72 terms per book (Sd. 1.13, std error mean 0.20.); while the indexers of CALIS totally assign 9,183 terms, averaging 2.78 terms per book (Sd. 1.27, std error mean 0.22). A t-test shows no significant difference between the two catalogues with regard to the number of index terms. That is to say, the two catalogues have identical level of indexing exhaustivity.

**Figure 1.** Consistency and total number of terms assigned by two indexers per book

The data shows whether two indexers assign equal number of index terms to the same book profoundly influences indexing consistency. With reference to Figure 1 we can see that when the total number of index terms assigned by two indexers to the same book is an even number, the indexing consistency is considerably high. In contrast, when the total number is an odd number, the indexing consistency is obviously low. Therefore, we should discuss the relationship between consistency and exhaustivity under two different conditions. Say condition 1 stands for the situation when the total number of terms assigned by two indexers to the same book is an odd number; condition 2 stands for the situation when the total number of terms assigned by two indexers to the same book is an even number.

Under condition 1, there is a positive correlation between consistency and exhaustivity. It indicates that the more terms two indexers totally assign to a book, the more likely they assign the same terms. Under condition 2, the relationship between consistency and the total number of terms assigned is a little more complicated. The consistency reaches a peak of 93.05\% when two terms are totally assigned to a book by two indexers, then it
declines sharply when four terms are assigned. But, after that it rises gradually, although there is a slight drop at the level of ten terms (see Figure 1). That is to say, in both cases consistency rises as the total number of terms assigned by two indexers per book increases, except for the extreme value 93.05%, which can be regarded as an outlier. In conclusion, consistency tends to be higher, when more terms are assigned to a book by two indexers.

Analysis

The experimental data has shown that when two indexers have identical level of exhaustivity, the indexing consistency is considerably high. Furthermore, the smaller differences in the number of terms assigned by two indexers lead to a greater probability of achieving higher consistency (see Figure 2). The reason is that when two indexers assign unequal number of terms to the same book (the total number is an odd number), there is no probability of achieving 100% consistency. Contrarily, when two indexers assign equal number of terms (the total number is an even number), there is a great probability of reaching 100% consistency.

Figure 2. The smaller differences in the number of terms assigned by two indexers, the higher consistency can be achieved

Actually, it is inevitable with the use of the two formulae mentioned above, because the formulae used increase the probability of a high score if two indexers assign an equal number of terms to a document rather than an odd number. It can be clearly explained by the following examples. In these examples, the consistency scores are calculated based on Hooper’s formula.

Example 1: If indexer A assigns 10 terms to a document, while indexer B assigns 10 terms to this document as well; then the number of commonly assigned terms $c$ can be 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. According to Hooper’s formula, the consistency scores can be 0, 1/19, 2/18, 3/17, 4/16, 5/15, 6/14, 7/13, 8/12, 9/11, 10/10. Assume that it is a discrete uniform distribution, and then the probability of each possible outcome is 1/11. The expected value of the consistency can be calculated by using the following function:

$$E(\text{consistency}) = \sum_{c=0}^{10} \frac{1}{11} \times \frac{c}{20-c}$$

In this case, the expected value is 0.40.
Example 2: If indexer A assigns 5 terms to a document, while indexer B assigns 15 terms; then the number of commonly assigned terms $c$ can be 0, 1, 2, 3, 4, 5. The consistency scores can be 0, 1/19, 2/18, 3/17, 4/16, 5/15. In this case, the expected value of the consistency is: 0.15.

Example 3: If indexer A assigns 1 term to a document, while indexer B assigns 19 terms; then the number of commonly assigned terms $c$ can only be 0 or 1. The consistency scores can be 0 or 1/19. In this case, the expected value of the consistency is: 0.03.

It is very clear that if the total number of terms assigned by indexer A and B is constant ($a+b$ is constant), the possibility that the two indexers are consistent with each other is the highest, when they assign equal number of terms. The smaller differences are between the numbers of terms they assign, the greater probability they get to achieve high consistency.

**Figure 3.** Consistency decreases with the increase of the total number of terms assigned to a book by two indexers, when the number of terms commonly assigned is constant.

Another interesting phenomenon is that when the number of terms commonly assigned to a book by two indexers is constant ($c$ is constant), the more terms the two indexers totally assign to this book, and the less consistent they are, which can be shown by the curves in Figure 3. As the total number of terms increases, the distances between these curves become smaller. That is to say, the more terms the two indexers totally assign to a book, the smaller are the differences in the consistency scores, no matter how many terms they commonly assign.

Figure 3 seems to contradict the conclusion that consistency tends to be higher, when more terms are assigned to a book by two indexers. It is essential to bear in mind that the curves in Figure 3 show the relationship between consistency and total number of terms assigned by two indexers per book, when the number of terms commonly assigned by two indexers is constant. However, in practice, the number of terms commonly assigned by two indexers is not constant. In fact, it varies with the total number of terms assigned by two indexers. The experimental data show that the more terms two indexers totally assign to a book, the more terms they commonly assign. Theoretically, it is also true, because when the total number of terms assigned by two indexers increases, the probability that they use the same terms is larger. In the light of Hooper’s formula, $c = \{0, 1,$
2, \ldots, (a+b)/2\}, where \(a+b\) is even; or \(c=\{0, 1, 2, \ldots, (a+b-1)/2\}\), where \((a+b)\) is odd. If we calculate the expected value of \(c\), we can find that the larger \(a+b\) is, the larger the expected value of \(c\) is.

The experimental data also show that the more common terms two indexers assign to a book, the more consistent they are (see Figure 4). However, it is not a linear relationship between the number of terms commonly assigned and indexing consistency. As the number of terms commonly assigned increases, indexing consistency infinitely approaches 100%, but will never reach it.

![Figure 4. Relationship between the number of terms commonly assigned to a book by two indexers and indexing consistency](image)

Since the more terms two indexers assign to a book, the more terms are in common, and meanwhile the more terms are in common, the higher consistency can be achieved, we can say that it is inevitable that two indexers are more consistent with each other when they assign more terms to a book.

**Discussion**

There is a common accepted statement about the relationship between indexing consistency and exhaustivity in this field: consistency drops as indexing exhaustivity increases. This belief appears to be supported by Fried & Prevel (1966, reported by Leonard 1977), Reich & Biever (1991), Sievert (1991) and Shoham (2001). However, this is not always the case, because the findings of the studies conducted by Harris (1966, reported by Leonard 1977), Leonard (1975, reported by Leonard 1977) contradict it.

In Reich and Biever’s (1991) study, they found that for a sample of articles indexed with an average of 8–9 thesaurus terms, the consistency was 24%; it was 45% for a sample having an average of 5–6 thesaurus terms. However, if we examine Reich and Biever’s results closely, we can find that it is highly debatable to draw such a conclusion based on their data, because about 16.67% of the descriptors were identical to terms in title in relation to sample #1, while about 46.55% in relation to sample #2. Theoretically, when indexers extract index terms directly from titles, they can be more consistent. Thus, the consistency was to some extent determined by how many descriptors were identical to terms in title. In this case, we cannot be sure whether the difference in consistency between sample #1 and sample #2 was resulted by different levels of exhaustivity or by different numbers of descriptors which were identical to terms in title.
Sievert (1991) compared the consistency of main headings (50.39%) with that of subheadings (47.89%), as well as comparing the consistency of descriptors (47.27%) with that of identifiers (32.83%). And he believed that the differences were resulted by the different levels of indexing depth. However, the properties of these different categories of index terms haven’t been taken into account in his research. Generally, main headings are used to represent major topics of a document, while subheadings are used to represent minor topics or specific aspects of a topic. Some evidence in the literature showed that indexers were more likely to agree on central aspects of a work than on peripheral themes (Lancaster, 1998; Funk et al., 1983; Iivonen, 1990). Besides, according to Sievert, in his sample, the descriptors were assigned from a small controlled vocabulary, while the identifiers were assigned from natural language. As it is known, controlled indexing has a beneficial effect on indexing consistency (Markey, 1984). Therefore, it is not convincing to say that higher exhaustivity results in lower consistency based on the findings of Sievert’s study.

Shoham and Kedar (2001) have made a considerable improvement over previous studies. They directly calculated the correlation between indexing consistency and number of terms, and found that the Pearson correlation coefficient was between -0.28 and -0.36 in different cases. Their findings indicated that there was a negative correlation between consistency and exhaustivity, although the strength of the association between consistency and number of terms was not substantially strong.

As mentioned above, some researchers took a contrary position. Harris (1966) maintained that ‘consistency (% overlap) does not seem to change much with depth… the graph (% consistency vs. depth) shows an almost horizontal line.’ Leonard (1977) also claimed that depth of indexing had no apparent effect on indexing consistency. Actually, whether two indexers assign equal number of index terms to the same book, i.e., whether they have identical indexing depth has profound effect on indexing consistency. Some evidence can be found in the literature.

Reich and Biever (1991) realized that different indexing depths could result in low consistency by saying: “the difference in the number of terms assigned to each document obviously contributes to the observed low index match rate of 27 percent.” If we examine Tonta’s study (1991), we can find the same reason for the low consistency (16% for exact matches and 36% for both exact and partial matches). According to his report, LC cataloguers assigned 282 subject headings for 82 items while BL cataloguers assigned 127. The difference in indexing depth between the two catalogues was significant. Stubbs’ experiment (1999) also provides some evidence. In the experiment, seven students indexed two printed manuals on library and information science twice. At the first indexing, indexer a, b, c and d were very selective (2–3 terms), indexer g was very exhaustive (30 terms), and indexer e and f applied 8 and 7 terms, respectively. But, at the second indexing, they had almost the identical exhaustivity (a 21, b 18, c 21, d 17, e 21, f 22, and g 25). Then, the mean consistency increased from 29% to 60%. There might be some other reasons for the increase. Nonetheless, we cannot deny that identical exhaustivity also contributed to the increase in consistency.

**Conclusion**

To sum up, the relationship between indexing consistency and exhaustivity is complicated. Whether two indexers assign equal number of index terms to a document affects profoundly indexing consistency. Although identical exhaustivity is not a sufficient condition
for high consistency, it is a necessary one. The analysis based on the experimental data of this study has shown that the widely accepted belief is actually wrong. There is no doubt that exhaustive indexing is beneficial rather than detrimental to indexing consistency. In the new networked environment, new optical level of indexing exhaustivity needs to be determined. Moreover, further studies are needed to understand the practical meanings of various consistency measures. Maybe a new consistency measure is needed, so that we can discuss the relationship between indexing consistency and exhaustivity without the influence of the measures used.

References
4W Vocabulary Mapping Across Diverse Reference Genres

Abstract
This paper examines three themes in the design of search support services: linking different genres of reference resources (e.g., bibliographies, biographical dictionaries, catalogs, encyclopedias, place name gazetteers); the division of vocabularies by facet (e.g., What, Where, When, and Who); and mapping between both similar and dissimilar vocabularies. Different vocabularies within a facet can be used in conjunction, e.g., a place name combined with spatial coordinates for Where. In practice, vocabularies of different facets are used in combination in the representation or description of complex topics. Rich opportunities arise from mapping across vocabularies of dissimilar reference genres to recreate the amenities of a reference library. In a network environment, in which vocabulary control cannot be imposed, semantic correspondence across diverse vocabularies is a challenge and an opportunity.

Introduction
The move away from cards and print to an online graphical user interface has important consequences for search support. This paper draws on studies at Berkeley to examine three different aspects of search support: overcoming the traditional separation between different genres of reference resource; the dividing of vocabularies by facet (e.g., What, Where, When, and Who); and mapping between different and dissimilar vocabularies. Important relationships between these three themes are noted.

Our starting point is the principle that understanding requires a knowledge of context and our approach to providing learners with contextual knowledge is to try to recreate the functionality of a traditional reference library in a digital environment.

Different Genres of Reference Resource
Several different forms of reference work have evolved for specialized purposes: Bibliographies, biographical dictionaries, chronologies, encyclopedias, library catalogs, place name gazetteers, and others. In a print environment any entry in any of these resources might cite any other, but, except for the library catalog, rarely tells one where a copy can be found. With the technology of card and paper these resources are not physically connected.

Sixty years ago, Ray Swank (1951) tried to combine the subject indexing of bibliographies with the locations listed in library catalogs, but this was not practical until the 1980s when digital technology became sufficiently available and affordable for records in published bibliographies to be linked to call numbers in local library records (Buckland 1988).

A digital, networked environment invites connectivity. Authority records on cards mention authoritative sources; authority records online can connect in real time with authoritative, explanatory resources, e.g., a place name authority record can link to a place name gazetteer. A link to a specialized reference work is likely to provide more and more up-to-date information than would be provided in an authority record. A well-formed place name gazetteer will not only couple each place name with spatial coordinates (latitude and longitude) for that place, but will also include a geographic description code (“feature type”) indicating what category of place it is (lake, city, mountain, etc.). In turn, the gazetteer could contain links to more extensive descriptions of the place in encyclopedias, pictures of it, map displays, socio-economic data series, and mentions in literary works.

We may think of representation in a knowledge organization system as different from description (as in an encyclopedia), but that distinction is not helpful here. The important
point is that the historic, physical separation of the catalog from reference works need not and should not continue.

**Divided Vocabularies**
A general purpose knowledge organization system, such as a list of subject headings or a library classification, will include all kinds of topics: activities, animals, events, institutions, persons, places, and so on. However, it can be more effective and probably more efficient to separate out categories of headings for separate treatment. In divided library card catalogs, personal names for the subjects of biographies are ordinarily separated from other subject headings and interfiled with the records for authors to constitute a name catalog for persons, both as authors and as subjects.

A graphical user interface allows the display of different kinds of relationships. For example, who, now, would want an online catalog without a map display showing where each place is and its geographical relationship to other places? This is feasible if the name of the place can be combined with its spatial coordinates of latitude and longitude. (Buckland, Chen, Gey, Larson, Mostern and Petras 2007).

We have been working with the four facets of WHAT (topic), WHERE (place), WHEN (time), and WHO (person) because these facets are different in kind and each has distinctive characteristics and display requirements: Semantic syndetic structure for topics; map displays and complex spatial relationships for places (e.g. Larson & Frontiera 2004); and time-lines and chronologies for time periods. Displays of family trees and other interpersonal relationships are desirable in biographies. Geotemporal arrangement of biographies into “life paths” or collective analysis as prosopography may be useful. There are other possibilities: We are currently considering the separation of named events from other topics (Lancaster, Buckland and Shaw 2007).

**Divided Vocabularies Combined in Representations**
Facets are, in principle, different in their nature, but topics often cannot be expressed in terms of a single facet. For example, 005.912=112.2(075)(410) representing “German language primers on office management in the U.K.” in the Universal Decimal Classification, is composed of a term from each of four facet vocabularies. Library of Congress Subject Headings (LCSH) routinely occur as compound representations in which the main heading is qualified by chronological and/or geographical subdivisions (e.g. Architecture—Japan—Meiji period, 1868-1912). (For discussion of LCSH and faceted interfaces see McGrath (2007)). Of course, textual descriptions in encyclopedia articles or elsewhere will use all kinds of vocabulary.

As already noted a well-formed place name gazetteer will not only couple each place name with the latitude and longitude for that place, but will also included a geographic description code (commonly called “feature type”) indicating what category of place it is (lake, city, mountain, etc.). Thus each WHERE is also categorized as a WHAT and, since names and boundaries are often unstable, should also be encoded for WHEN.

The treatment of time is less well-developed than place. In ordinary discourse and in metadata it is common to express time in geographical and cultural terms by using named events adjectivally to denote period, e.g. Victorian literature, a Civil War weapon, Louis Quinze furniture, and so on. LCSH chronological subdivisions ordinarily define time using periods named in terms of the political history of the topic (e.g. « Meiji period ») with calendar dates added for clarification. The use of named events to denote spans of time
has more evocative connotations than mere calendar time because a cultural context is identified that might otherwise not have been recognized when relying on dates alone. Calendar time is needed, however, to map temporally between the periods of different cultures. (Petras, Larson and Buckland 2006).

Biographical texts are rich in actions, locations, dates, and other persons, but mark up standards and best practices are still generally unsatisfactory (Text Encoding Initiative 2006). We are examining the feasibility of decomposing lives into a series of events and using 4-tuple of What, Where, When, and Who to characterize each life event: an activity or event (birth, education, employment, etc.) in a place at some time, sometimes with other people (Electronic Cultural Atlas Initiative 2006).

The use of terms from different facets in compound representations and textual descriptions provides an opportunity to link any identifiable What, Where, When, or Who to an explanatory description in some other reference work. The examples given above can be briefly summarized:

Table 1. Components of examples different reference genres

| WHAT (LCSH): | Topic — Geographical subdivision — Chronological subdivision |
| WHERE (Gazetteer): | Place name — Feature type — Latitude and longitude — When |
| WHEN (Chronology): | Period name — Period type — Dates — Where |
| WHO (Biogr. Dict.): | Personal name — Actions — Places — Dates — Other persons |

These examples provide the basis for a two-dimensional 4W array: The WHAT records have WHERE and WHEN subdivisions; the WHERE records (place name gazetteer entries) should have WHAT (feature type) and WHEN elements; WHEN records in our design for a time period directory include elements for WHAT (period type) and WHERE; and a WHO record in a biographical directory will contain multiple WHAT (action, status), WHERE, WHEN, and WHO elements. We can note a WHAT (Topic, Type, Type, Action) element in each row and also WHERE and WHEN in each. Rearranging the components in each row yields a new table in which the facets are aligned vertically.

Table 2. Two dimensional array of components of reference genres

| WHAT (LCSH): | What | Where | When |
| WHERE (Gazetteer): | What | Where | When |
| WHEN (Chronology): | What | Where | When |
| WHO (Biogr. Dict.): | What | Where | When | Who |

Displayed this way, the opportunities for mapping vertically between vocabularies in reference genres become apparent. Semantic associations can be established by vertical vocabulary mapping and contextual associations can be established through horizontal associations.

Mapping Between Similar Vocabularies
Ordinarily vocabulary control is within a single vocabulary and interoperability through vocabulary mapping is between resources of a similar type, meaning at the same horizontal level in Table 1. Similar resources commonly use different knowledge organization vocabularies that are the same (or overlap) in scope but differ in form. The Library of Congress Subject Headings (LCSH), the Dewey Decimal Classification and the Library of Congress Classification are different and look different (e.g. for Economics: “Economics,” “330,” and “HB1”), yet they are similar in that they are general purpose systems for providing top-
ical access. Vocabulary interoperability is ordinarily based on cross-walks defining comparable fields in two or more similar resources and mapping between terms in those fields.

**Mapping Between Dissimilar Vocabularies**

The attraction of linking between *different genres* of reference resource is the wider range of descriptions and so the richer context. The vertical columns Table 2 reveal a greatly expanded scope for mapping between dissimilar vocabularies. For example, the feature type codes of a place name gazetteer can be mapped to the topics in a subject heading list. We examined the 600+ Geographic Description Codes (GDC) of the US federal National Geointelligence Agency in relation to the far larger Library of Congress Subject Headings and found that, despite stylistic differences, there was, in most cases, an identifiable equivalent. This mapping allows a connection between a category of geographical feature (e.g. a lighthouse) and literature about lighthouses. In the other direction one could go from literature to physical examples, their locations, and a map display showing where they are located. One could move, for example, from the Geographic Description Code “school” to the corresponding library catalog subject heading and on to related literature. Since a gazetteer is concerned with physical, geographical features the semantic equivalent of the geographical feature « School » in LCSH is “School buildings,” but, situationally, “Schools” (denoting schools as institutions) may be preferred. In both GDC and LCSH the option of displaying related terms would be highly desirable (Buckland, Chen, Gey, Larson, Mortern and Petras 2007). For examples of literature being related to maps see Moretti (1988).

Such mapping may seem inappropriate because subject headings are concerned with topics and feature types with physical objects and these are different in kind and, therefore, not comparable. This criticism is correct in principle, but misguided. It is precisely this difference that enables one to move from literature to physical objects and vice versa, a genuinely multimedia approach.

**Why Vocabulary Mapping Is Hard**

There is a widely-held assumption that search support in the Web should be provided by ontologies and that these ontologies can be made interoperable. (DeRidder 2007 provides an overview). This seems reasonable until one tries to do it and/or one reflects on the nature of language and/or is concerned with representations of knowledge rather than of widgets and/or contemplates the size of the task. Mapping between knowledge organization vocabularies is hard for several reasons:

1. Vocabularies are descriptive. Assigning descriptive metadata (descriptors, index terms, classification codes, category codes) is a language activity even through documentary languages are more or less artificial. They are incorrigibly context-specific and inherently obsolescent (Buckland 2007).
2. Vocabularies are cultural. The classic definition by E. B. Tylor (1871, v. 1, p. 1) states “Culture or Civilization, taken in its wide ethnographic sense, is that complex whole which includes knowledge, belief, art, morals, law, custom, and any other capabilities and habits acquired by man as a member of society.” More recent definitions refer to integrated patterns of human knowledge, belief, and behavior, agree that culture is not genetically transmitted but learned, and position knowledge as a part of culture. It follows that knowledge organization is concerned with learned, cultural entities and it is a characteristic of what is cultural (and of what is learned) that it is subjective, more or less imprecise, and unstable. Discourse about
culture tends to use archetypes for lack of clear definitions and instances often form a smooth continuum from one type to another. The implications for knowledge organization are significant. In general terms, it must mean that precise, logical representations are unavoidably Procrustean distortions. Vocabularies such as thesauri and classification schemes are valuable and necessary. Nevertheless vocabularies become increasingly problematic as their use is extended beyond a single, local application and/or are continued over time.

3. Language, being cultural, evolves within domains of discourse and it is not to be expected that any given vocabulary will correspond precisely to one homogenous community. Typically a knowledge organization vocabulary is a linguistic compromise. Ideally, there would multiple indexes to any given resource, one for each community of users, a challenge addressed by Petras (2006).

4. There is a lot mapping to be done! The whole logic of an internet environment is to greatly increase access to resources with more-or-less unfamiliar descriptive metadata, greatly increasing the need and scope for mapping between vocabularies. Mapping is best done by well-qualified experts, but it is very labor-intensive. Fortunately, statistical association techniques combined with natural language processing can provide rapid, inexpensive search term recommender services where a training set can be found (Buckland and others 1999; Gey, M. Buckland, A. Chen and R. Larson. 2001).

Conclusions
Understanding requires a knowledge of context so it is important to enable learners to use diverse reference genres to find contextual information. In a traditional, paper-based reference library the various resources are not physically linked but cite each other and are physically collocated so that it is humanly possible to move from an entry in one to an entry in another. That functionality has yet to be reconstructed in a digital library environment (Buckland 2008), but the adoption of online graphical user interfaces allows many important developments. Dividing vocabularies by facet simplifies the task of assigning descriptive metadata, facilitates specialized displays, and constitutes a form of infrastructure (Buckland 2006). Combining divided vocabularies is necessary for representation and descriptions. Mapping across both similar and dissimilar vocabularies is a difficult but necessary part of a broadly based learning environment.

Acknowledgments
This paper draws on the work of several other people, especially Fredric Gey, Ray R. Larson and Vivien Petras. We are grateful for the support of the Institute of Museum and Library Services National Leadership Grant for Libraries, LG-02-02-0035-02; LG-02-04-0041-04; and LG-06-06-0037-06 and to both the National Endowment for the Humanities and the Institute of Museum and Library Services for an Advancing Knowledge grant PK-50027-07 (Electronic Cultural Atlas Initiative. 2002, 2004, 2006 and 2007).

References


A Survey of the Top-level Categories in the Structure of Corporate Websites

Abstract
A survey of the structure of corporate Websites was carried out to identify common categories, structures, facets and terms used in organizing these Websites. A sample of seven Websites of IT companies were first analyzed to construct a taxonomy checklist with 82 categories. The checklist was then used to analyze the top-level structure of 28 corporate Websites belonging to six product types. Differences in the top-level categories were found across the product types. Fifty-four new categories were found and incorporated into the taxonomy checklist. The taxonomy checklist is expected to be a useful reference for Information Architects in the design of corporate Websites.

Introduction
Corporate and public sector organizations are increasingly using infrastructure services such as Websites, intranets, portals and institutional repositories to leverage knowledge resources in the organization by their employees and customers. Studies, however, have shown that discovering relevant knowledge resources on these sites is becoming problematic because of lack of proper content organization (Feldman & Sherman, 2001; Linden, 2001). Various papers have emphasized that improved content organization, particularly by deploying classifications schemes, controlled vocabularies, taxonomies, and metadata, can enhance the resource discovery and information findability (Rowley, 2000; Duffy, 2001; Delphi Group, 2002). While reviews of features of intranets and Websites have been undertaken by various researchers (e.g., Hoffman, 1998; Watson & Fenner, 2000; Karayanni & Baltas, 2003; Cappel & Huang, 2007), very few have focused on organization of contents and related issues.

We have embarked on a project to study organizational taxonomies used for organizing Websites, enterprise portals and institutional repositories. This paper reports the initial results of a survey of corporate Websites carried out to identify common structures, facets, categories and terms used in organizing these Websites. This survey is limited to Websites of multinational companies that sell products. The Websites therefore serve the purpose of providing information about products to potential customers, as well as offer various kinds of online customer support to existing customers. Being multinational and serving a global customer base, the Websites tend to be quite large and well-developed, and provide a wealth of information about the company and its products and services.

The study assumes that the organization of a Website can be represented by taxonomy of concepts and terms. Indeed it is assumed that designing the information architecture of a Website involves a first step of constructing a taxonomy as an abstract representation of the structure of the Website and organization of its contents. The taxonomy is then expressed as a navigation or search structure, manifested in one of many possible menu designs or interaction designs, and instantiated in a visual design (graphics design). Adopting the framework of the Functional Requirements for Bibliographic Records (IFLA Study Group, 1998), we view the organizational taxonomy as a Work (intellectual creation), the navigation or search structure as an Expression of the taxonomy, the menu system and interaction objects such as hyperlinks and wireframe as the Manifestation of
the navigation/search structure, and the Website incorporating the graphics or visual design as an Item or Exemplar of the Manifestation.

The survey of corporate Websites is implemented in the following three phases:
1. Analyzing a small sample of Websites to identify common facets, common categories in each facet, and the common structure of each facet;
2. The common facets, categories and structures are then used as a taxonomy checklist to analyze the Websites that are surveyed in our study.
3. The metadata elements used to describe knowledge resources on the Websites are analyzed.

The taxonomy checklist thus serves as a data collection instrument to study the structure of company Websites.

This paper reports the initial results of the analysis of the top-level structure of the Websites to identify the common categories and facets. The taxonomy checklist is then improved in the light of the survey results, and can be used as a reference by Information Architects when designing corporate Websites.

Previous Work
There have been a handful of studies on the use of taxonomies for Websites and portals. McDavid (1996) and Bailey (1994) suggested that orderly classification of terms and taxonomy was the key to success for reducing the shortfalls associated with inefficiency. Gilchrist (2001) reviewed corporate practices in content organization and the use of taxonomies within corporate information portals. He presented six case studies after a survey of 22 corporations. Schnebel (2002) conducted a case study of a knowledge portal at Cisco that used an Integrated Knowledge Architecture (IKA) approach to organize contents. IKA organizes information into smaller pieces called knowledge objects and builds meaningful relationships between these objects. Hudon and Hjartarson (2002) reviewed taxonomies and thesauri that were used by government sites in a number of countries. Renaud (2004) reviewed the standards and activities to create and manage metadata and controlled vocabularies in Canada, and reported a number of knowledge organization schemes used on government sites. Shir and Molberg (2005) reviewed the knowledge organization systems deployed by Canadian Government Web Sites (mainly digital collections). They reported that 33 digital collections made use of some sort of knowledge organization system—thesauri, subject headings, and classification schemes. They concluded that searching, browsing, and navigation facilities required improvements. Kremer, Kolbe and Brenner (2005) reported that an international professional services firm, with 150,000 employees in more than 150 countries worldwide, created a corporate taxonomy to organize the content of their portal. More recently, Wiggins, Remley, and Klinger (2006) described the creation of content management system for the Website of Kent State University Libraries and Media Services that used a schema to accommodate and manage extensive metadata descriptions, making it easier for users to find information needed from the Website.

Method
The survey was carried out from October 2007 to January 2008. The taxonomy checklist was developed based on a small sample of seven IT companies:
• IT products — hardware: IBM, HP, Nokia, Apple, Sun
• IT products — software: Microsoft, Oracle.
The categories found in the menu and navigation systems (including hyperlinks) of the seven Websites were first harvested from the home pages and all Web pages linked to the home pages. Categories that appeared in multiple Websites as well as those that were deemed by the researchers to be generic and not specific to a particular company were shortlisted. The shortlisted categories were organized into facets, and similar categories were merged to form the taxonomy checklist (listed in Appendix 1).

The taxonomy checklist was then used in a survey of 28 company Websites to find out how common the categories were. The survey also serves to validate the checklist. For the survey, four or five multinational companies were selected for each of the following product types:

- IT products-software: Sybase, Adobe, Symantec, SAP
- IT products-hardware: Dell, Intel, Toshiba, Samsung, Fujitsu
- Consumer electronics: Sony, Philips, Panasonic, LG
- Automobile: Rolls Royce, Ford, Chrysler, Toyota, General Motors
- Furniture: IKEA, Thomasville, Vitsoe, HNI Corporation, Artek

These product types are familiar to almost every household, and the products have a wide customer base.

For each Website, menu terms and hyperlinked terms were extracted from the home page and all Web pages linked to the home page. However, if the number of pages linked to the home page is fewer than 5, then the third level of Web pages was also processed.

Results

For the 82 categories in the taxonomy checklist, the frequencies of occurrence in the 28 company Websites are listed in Appendix 1. The frequency of occurrence refers to the number of Websites having the category at the top levels of the navigation system (mainly level 1 and 2). The frequencies listed in Appendix 1 are subdivided by product type and level 1 and 2. The statistics for level 2 includes a small number of level 3 and 4 categories. This is because, for the smaller Websites, we also analyzed the level 3 Web pages.

All the categories in the checklist were found to be fairly common in the corporate Websites that we analyzed, each category occurring in at least six Websites. Thirty-six categories were found to occur in more than half the Websites. Furthermore, the top category of each facet in the taxonomy checklist tend to have higher frequencies, providing some validation for the facets.

Since the companies sell products and services to consumers, many of the common categories relate to products, services, customer support and online shopping. The second set of common categories are generic categories that are likely to be present in all corporate Websites: contact information, company profile and news. What is surprising perhaps is the prevalence of the category for investors. The companies also appear to be aware of the importance of cultivating their customer online communities.

The frequency table in Appendix 1 shows differences in category frequencies across the different product types. Hardware, software and automobile Websites tend to have more categories referring to product information (e.g. documents, white papers and descriptions), whereas consumer electronics, furniture and household products sites appear to carry little product information, at least at the top levels of the Websites. Software and demos relevant only to IT product sites. Customer support information, support resources
and online communities also seem to be unimportant to furniture and household products Websites. Perhaps for these product types, customers tend to obtain product information directly from the stores, rather than seek them online.

The differences between the product types appear to be related to the cost and complexity of the product. For the more expensive and complex products such as IT products and automobiles, more information is sought by customers online. IT product customers are also likely to be more IT savvy and more likely to participate in online communities. Not surprisingly, the taxonomy checklist fits IT product Websites better than the Websites for other product types.

Fifty-four new categories were identified in the 28 Websites, listed in Appendix 2. The new categories are mainly in the areas of careers and jobs, product information, and more information about the company, including corporate culture, community relations and financial performance. Companies appear to be taking pains to project a good corporate image via their Websites. A new version of the taxonomy checklist incorporating the new categories is being constructed. The taxonomy will then be validated against a new sample of Websites.

Further analysis of the category labels used in the Websites and the navigation structure is in progress. The metadata attached to the Web pages are also being analyzed to find out whether metadata is used systematically to describe the resources, what are the common metadata elements across companies, and any standard metadata schemas used.

Future Work
The second phase of the project will comprise follow-up questionnaire and interview surveys aimed at validating information collected through the survey of Websites. Additional information (on policies, procedures, and related issues) that is difficult to extract from Websites will be collected from relevant officers of the organizations.

Follow-up user studies can be carried out to investigate the effectiveness of the different facets in the taxonomy checklist, how they can be improved, and how they should be used in the Website navigation system. The study will be also extended to cover government Websites, publicly-funded organizations and e-citizen portals.

The results of this project are expected to be of interest to organizations interested in development of tools and techniques for organizing contents on organizational sites and repositories. The findings may be helpful in identifying common patterns that can be used to recommend guidelines for improved content organization strategies.

References


### Appendix 1. List of categories and number of company Websites having the category, for six product types

<table>
<thead>
<tr>
<th>Product Type</th>
<th>Software</th>
<th>Hardware</th>
<th>Automobile</th>
<th>Electronics</th>
<th>Furniture</th>
<th>Household</th>
<th>Total (N=28)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level</strong></td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1 &amp; 2</td>
</tr>
<tr>
<td><strong>Products</strong></td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Documents</strong></td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>White papers</strong></td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Hardware</strong></td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td><strong>Software</strong></td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Trials/Demos</strong></td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Services</strong></td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Business services</strong></td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>Consulting services</strong></td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>IT services</strong></td>
<td>0</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Support</strong></td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td><strong>Certification</strong></td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>Order support</strong></td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td><strong>Product support</strong></td>
<td>0</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td><strong>Recycle/Resale</strong></td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Repair</strong></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>Service programs</strong></td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td><strong>Services support</strong></td>
<td>0</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td><strong>Support policies</strong></td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Training</strong></td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>Warranty</strong></td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Contact us</strong></td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td><strong>Email</strong></td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td><strong>Locations</strong></td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Office</strong></td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Reseller</strong></td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>Service centre</strong></td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td><strong>Worldwide locations</strong></td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Online chat</strong></td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Phone/Fax</strong></td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Service request</strong></td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Support resources</strong></td>
<td>0</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Communities</strong></td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>FAQ</strong></td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td><strong>Library/Support database</strong></td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td><strong>Product manuals</strong></td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td><strong>Publications</strong></td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Support news</strong></td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Troubleshoot</strong></td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Downloads</strong></td>
<td>0</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td><strong>Developer tools</strong></td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Drivers</strong></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Fixes/Patches</strong></td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>New downloads</strong></td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Top downloads</strong></td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Updates</strong></td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>User groups</strong></td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Businesses</strong></td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Small businesses</strong></td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Midsize businesses</strong></td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Large businesses</strong></td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Developers</strong></td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Home users</strong></td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Investors</strong></td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td><strong>IT pros</strong></td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Organizations</strong></td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Partners</strong></td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>Press</strong></td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td><strong>About</strong></td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

<p>| Total (N=28) | 26 | 9 | 17 | 10 | 9 | 9 | 27 |</p>
<table>
<thead>
<tr>
<th>Product Type</th>
<th>Software</th>
<th>Hardware</th>
<th>Automobile</th>
<th>Electronics</th>
<th>Furniture</th>
<th>Household</th>
<th>Total (N=28)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>1 2 2 1 2 2 1 2</td>
<td>1 2 1 2 1 2 1 2</td>
<td>1 2 1 2 1 2 1 2</td>
<td>1 &amp; 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advertising</td>
<td>4 0 5 0 4 0 0 2</td>
<td>1 2 0 1 0 1 0 2</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Awards</td>
<td>2 2 1 4</td>
<td>2 2</td>
<td>0 1 0 1 0</td>
<td>2 0 2 0 2</td>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Company info</td>
<td>0 4 5 0</td>
<td>2 2</td>
<td>1 3 0 3 0</td>
<td>3 2 23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corporate governance</td>
<td>1 3 1 4</td>
<td>3 1</td>
<td>0 1 1 0</td>
<td>0 4 1 9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Directors</td>
<td>0 4 4 1</td>
<td>0 0</td>
<td>0 0 0 1</td>
<td>0 2 1 12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Executives</td>
<td>0 4 0 5</td>
<td>0 0</td>
<td>0 1 0 1</td>
<td>0 3 0 3 14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>History</td>
<td>0 4 0 5</td>
<td>1 3</td>
<td>0 1 0 3</td>
<td>0 3 0 5 22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organization structure</td>
<td>2 2 0 5</td>
<td>0 3</td>
<td>0 1 0 0</td>
<td>0 1 0 1 14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vision</td>
<td>4 0 2 3</td>
<td>2 2</td>
<td>0 0 2 0</td>
<td>2 0 2 17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portal/Login/My …</td>
<td>4 0 5 0</td>
<td>1 2</td>
<td>0 1 0 1</td>
<td>0 1 0 1 15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>News</td>
<td>3 1 5 0</td>
<td>4 0</td>
<td>1 2 2 0</td>
<td>3 2 2 23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Events</td>
<td>4 0 5 0</td>
<td>4 0</td>
<td>0 3 0 1</td>
<td>0 3 0 3 20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Web casts</td>
<td>4 0 5 0</td>
<td>4 0</td>
<td>0 0 0 1</td>
<td>0 1 0 4 18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Store/Shopping</td>
<td>4 0 5 0</td>
<td>2 1</td>
<td>0 3 0 1</td>
<td>0 1 0 1 16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Industry</td>
<td>3 1 5 0</td>
<td>3 0</td>
<td>1 0 0 0</td>
<td>0 0 0 0 13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finance</td>
<td>2 2 1 0</td>
<td>1 0</td>
<td>0 0 0 0</td>
<td>0 0 0 0 6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Government</td>
<td>2 2 3 0</td>
<td>0 0</td>
<td>0 0 0 0</td>
<td>0 0 0 0 7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health</td>
<td>2 2 2 0</td>
<td>0 0</td>
<td>0 0 0 0</td>
<td>0 0 0 0 6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Appendix 2. New categories found in the survey**

**Products**
- Brochure
- Build your own/
- Personalization
- Comparisons
- Design
- Engineering
- Experience
- Innovation
- New products
- Parts and accessories
- Technology
- What our customers say
- Why …
- Careers/Employment
- College recruiting
- Internships
- Scholarships
- Why …
- Choose a region

**About …**
- Annual Report
- Corporate culture
- Community projects
- Corporate causes
- Corporate governance
- Corporate
- responsibility
- Diversity
- Environment
- Financial performance
- Glossary
- Legal
- Life at …
- Picture gallery
- Privacy
- Safety & health
- Social responsibility
- Stock information
- Security
- Use and disclosure of information

**Contact us**
- Contact investor
- Contact media
- Donations
- Feedback
- Idea for new product
- Operating hours
- Service hotline
- Service
- Press releases
- News

**Search**
- Search by Products/
- Service
- Sitemap
- Store/Shopping
- Buy
- Financing
- Gift
- Promotions and offers
- Request a quote
- Shopping tools
- Trade-in value
- Owners
- Students
- Supplier
A Case Study of Tagging Patterns in del.icio.us

Abstract
This paper presents a longitudinal case study and analysis of tagging patterns in del.icio.us. Previous research has indicated that a tagging vocabulary will stabilize over time, suggesting that convergence may occur. This case study investigates the possibility of stability and convergence in a subset of the tagging vocabulary used with del.icio.us.

Introduction
In social tagging systems, each user creates a folksonomy or personalized collection of bookmarks by assigning keywords, or tags, to web resources. The user’s collection of tags reflects his personal categorization of each resource in a flat, non-hierarchical structure. Therefore, for the tagging vocabulary of a single user, the unit of analysis is the resource-tag pair.

Although such a collection and its associated folksonomy are necessarily tailored to the whims of the individual user, the fact that the tagging vocabulary may be shared across users makes it a convenient tool for searching, browsing, and retrieving resources. A social tagging system, such as that of del.icio.us or Flickr, is a collection of folksonomies created by individual users. In a social tagging system, a tag assigned for personal use may serve as a link to additional resources that have been tagged by other users; and, because a social tagging system can also indicate connections among users, the unit of analysis becomes the resource-tag-user triple. Since each triple implies a unique path in this network, there may be users who share the same resource-tag relation and resources that share the same user-tag relation. In this way, the personal tags in each folksonomy become a link between the categorization system of the individual and the collective categorization reflected in the representation of resources in the network.

As an organic, non-restricted network, a social tagging system suffers from variations in the resource-tag-user triple that reflect conflicting semantic relations between words and their referents (Golder & Huberman, 2006). There is systematic variation across individuals, whether in knowledge expertise, social factors or personal goals, that affect the formation of resource-tag pairs (Golder & Huberman, 2006; Marlow et al., 2006). Two recent studies of Del.icio.us also suggest that the pattern of resource-tag pair adopted by a user may be affected by imitation, knowledge sharing, or the robustness of cognitive concepts (Golder & Huberman, 2006; Monk & Mork, 2007a). In a widely cited study of tagging behaviors in Del.icio.us, Golder and Huberman observed stable patterns where each tag’s frequency of occurrence for one resource assumed an almost fixed proportion to the total frequency of all tags used for that resource after the first 100 resource-tag-user triples had been assigned. Such a distribution implies stability of the relationship between resources, tags, and users as the collective tagging system evolves over time, despite the variations in individual and social factors mentioned previously.
Similarly, Monk and Mork (2007b) have identified regularities in the way tags are used, in their frequency, and in the relative distribution of frequency between tags. In general, they indicated that user-generated tags moved toward consensus on assignment of specific tags to specific web resources over time. They theorized that this pattern follows a power-law distribution that occurs in small world complex networks: the number of user-tag pairs declines at an exponential rate as the number of resource-tag pair declines (Shen & Wu, 2005). Monk and Mork (2007b) further discovered that not only the number and proportion of tags stabilized over time, but that the meaning of those tags was also aggregated in a predictable pattern. In this power-law of distribution, broad and general content tags dominated, while the majority of tags were peripheral and rarely used.

Methodology
For a tagging vocabulary to be truly useful, it should exhibit both stability (i.e., regularity of application) and convergence (i.e., growing consensus regarding the assignment of tags). To gather insight as to whether or not stability and convergence can be identified in the set of tags assigned to a given URL, a purposive or judgement sample was collected from del.icio.us. Using *rdf* as a seed term, all URLs that had been tagged with *rdf* at least once were collected using del.icio.us RSS feeds from Google Reader caches. Because Google Reader archives the RSS feeds to which its users subscribe, it was possible to collect all URLs tagged with *rdf* between October 7, 2005, and January 30, 2008. Based on the RSS feeds, a list of 7845 unique URLs was created and data for these resources was downloaded from the del.icio.us site during the period from February 8 to 12, 2008. Based on the set of unique tags that made up 68.3% of the total number of assigned tags was then identified (see Table 2). These tags were analyzed and the individual tags were plotted for their frequency of occurrence across time. The threshold for identifying a set of tags for each of the ten URLs was arbitrarily set at 68.3% to approximate one standard deviation away from the mean.

Stability
In this study, stability was operationalized as the consistency of a tag’s proportion of the total vocabulary over time. For each tag in the set of unique tags that made up 68.3% of the total number of tags assigned to a given URL, a measure of tag stability was computed as the proportion of the tags’s use in a given month to the total number of tags assigned for that month. By plotting a tag’s percentage of use across time, it is possible to evaluate the tag’s stability as consistency of occurrence independent of the total number of tags assigned during any particular time period.
Table 1. Ten URLs most frequently tagged rdf

<table>
<thead>
<tr>
<th>URL</th>
<th>Total rdf tags</th>
<th>Total Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://simile.mit.edu/">http://simile.mit.edu/</a></td>
<td>559</td>
<td>10,563</td>
</tr>
<tr>
<td><a href="http://www.w3.org/RDF/">http://www.w3.org/RDF/</a></td>
<td>551</td>
<td>2,729</td>
</tr>
<tr>
<td><a href="http://jena.sourceforge.net/">http://jena.sourceforge.net/</a></td>
<td>426</td>
<td>3,287</td>
</tr>
<tr>
<td><a href="http://www.w3.org/TR/rdf-primer/">http://www.w3.org/TR/rdf-primer/</a></td>
<td>417</td>
<td>1,794</td>
</tr>
<tr>
<td><a href="http://www.foaf-project.org/">http://www.foaf-project.org/</a></td>
<td>387</td>
<td>5,646</td>
</tr>
<tr>
<td><a href="http://www.xml.com/pub/a/2001/01/24/rdf.html">http://www.xml.com/pub/a/2001/01/24/rdf.html</a></td>
<td>369</td>
<td>1,585</td>
</tr>
<tr>
<td><a href="http://librdf.org/">http://librdf.org/</a></td>
<td>355</td>
<td>1,813</td>
</tr>
<tr>
<td><a href="http://www.w3.org/TR/rdf-sparql-query/">http://www.w3.org/TR/rdf-sparql-query/</a></td>
<td>355</td>
<td>2,353</td>
</tr>
<tr>
<td><a href="http://protege.stanford.edu/">http://protege.stanford.edu/</a></td>
<td>347</td>
<td>4,712</td>
</tr>
<tr>
<td><a href="http://www.w3.org/2001/sw/">http://www.w3.org/2001/sw/</a></td>
<td>345</td>
<td>4,886</td>
</tr>
</tbody>
</table>

Table 2. Frequency of tags assigned to http://librdf.org/ and cumulative percent of total tags assigned. Tags in greyed cells were not included in the analysis

<table>
<thead>
<tr>
<th>Tag</th>
<th>Frequency</th>
<th>Percent</th>
<th>Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>rdf</td>
<td>355</td>
<td>19.6</td>
<td></td>
<td>19.6</td>
</tr>
<tr>
<td>semanticweb</td>
<td>148</td>
<td>8.2</td>
<td></td>
<td>27.7</td>
</tr>
<tr>
<td>programming</td>
<td>118</td>
<td>6.5</td>
<td></td>
<td>34.3</td>
</tr>
<tr>
<td>redland</td>
<td>110</td>
<td>6.1</td>
<td></td>
<td>40.3</td>
</tr>
<tr>
<td>library</td>
<td>98</td>
<td>5.4</td>
<td></td>
<td>45.7</td>
</tr>
<tr>
<td>xml</td>
<td>85</td>
<td>4.7</td>
<td></td>
<td>50.4</td>
</tr>
<tr>
<td>python</td>
<td>69</td>
<td>3.8</td>
<td></td>
<td>54.2</td>
</tr>
<tr>
<td>web</td>
<td>47</td>
<td>2.6</td>
<td></td>
<td>56.8</td>
</tr>
<tr>
<td>framework</td>
<td>40</td>
<td>2.2</td>
<td></td>
<td>59.0</td>
</tr>
<tr>
<td>semantic</td>
<td>36</td>
<td>2.0</td>
<td></td>
<td>61.0</td>
</tr>
<tr>
<td>software</td>
<td>35</td>
<td>1.9</td>
<td></td>
<td>62.9</td>
</tr>
<tr>
<td>opensource</td>
<td>34</td>
<td>1.9</td>
<td></td>
<td>64.8</td>
</tr>
<tr>
<td>api</td>
<td>32</td>
<td>1.8</td>
<td></td>
<td>66.6</td>
</tr>
<tr>
<td>development</td>
<td>32</td>
<td>1.8</td>
<td></td>
<td>68.3</td>
</tr>
<tr>
<td>ruby</td>
<td>32</td>
<td>1.8</td>
<td></td>
<td>70.1</td>
</tr>
<tr>
<td>c</td>
<td>29</td>
<td>1.6</td>
<td></td>
<td>71.7</td>
</tr>
<tr>
<td>tools</td>
<td>26</td>
<td>1.4</td>
<td></td>
<td>73.1</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>Total</td>
<td>1,813</td>
<td>100.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Percent of tag occurrence by month for http://www.w3c.org/RDF

<table>
<thead>
<tr>
<th>Month</th>
<th>rdf</th>
<th>w3c</th>
<th>semanticweb</th>
<th>xml</th>
<th>standards</th>
<th>web</th>
<th>semantic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan-07</td>
<td>17.24%</td>
<td>10.34%</td>
<td>6.90%</td>
<td>8.05%</td>
<td>9.20%</td>
<td>3.45%</td>
<td>5.75%</td>
</tr>
<tr>
<td>Feb-07</td>
<td>17.58%</td>
<td>9.89%</td>
<td>7.69%</td>
<td>4.40%</td>
<td>7.69%</td>
<td>6.59%</td>
<td>5.49%</td>
</tr>
<tr>
<td>Mar-07</td>
<td>18.27%</td>
<td>9.62%</td>
<td>13.46%</td>
<td>6.73%</td>
<td>4.81%</td>
<td>4.81%</td>
<td>4.81%</td>
</tr>
<tr>
<td>Apr-07</td>
<td>22.39%</td>
<td>7.46%</td>
<td>4.48%</td>
<td>8.96%</td>
<td>5.97%</td>
<td>7.46%</td>
<td>5.97%</td>
</tr>
<tr>
<td>May-07</td>
<td>20.19%</td>
<td>10.58%</td>
<td>12.50%</td>
<td>9.62%</td>
<td>7.69%</td>
<td>0.96%</td>
<td>5.77%</td>
</tr>
<tr>
<td>June-07</td>
<td>19.23%</td>
<td>9.62%</td>
<td>10.58%</td>
<td>6.73%</td>
<td>9.62%</td>
<td>7.69%</td>
<td>2.88%</td>
</tr>
<tr>
<td>July-07</td>
<td>20.79%</td>
<td>11.88%</td>
<td>6.93%</td>
<td>8.91%</td>
<td>8.91%</td>
<td>7.92%</td>
<td>6.93%</td>
</tr>
<tr>
<td>Aug-07</td>
<td>19.15%</td>
<td>6.38%</td>
<td>12.77%</td>
<td>6.38%</td>
<td>4.26%</td>
<td>4.26%</td>
<td>6.38%</td>
</tr>
<tr>
<td>Sep-07</td>
<td>23.44%</td>
<td>10.94%</td>
<td>15.63%</td>
<td>6.25%</td>
<td>6.25%</td>
<td>1.56%</td>
<td>3.13%</td>
</tr>
<tr>
<td>Oct-07</td>
<td>16.67%</td>
<td>9.65%</td>
<td>11.40%</td>
<td>7.89%</td>
<td>8.77%</td>
<td>5.26%</td>
<td>4.39%</td>
</tr>
<tr>
<td>Nov-07</td>
<td>15.22%</td>
<td>5.43%</td>
<td>8.70%</td>
<td>8.70%</td>
<td>5.43%</td>
<td>7.61%</td>
<td>8.70%</td>
</tr>
<tr>
<td>Dec-07</td>
<td>22.06%</td>
<td>8.82%</td>
<td>8.82%</td>
<td>2.94%</td>
<td>7.35%</td>
<td>4.41%</td>
<td>1.47%</td>
</tr>
<tr>
<td>Jan-08</td>
<td>22.22%</td>
<td>9.26%</td>
<td>12.96%</td>
<td>3.70%</td>
<td>7.41%</td>
<td>7.41%</td>
<td>3.70%</td>
</tr>
</tbody>
</table>
Table 4. Percent of tag occurrence by month for http://libRDF.org

<table>
<thead>
<tr>
<th>Tag</th>
<th>Aug-07</th>
<th>Sep-07</th>
<th>Oct-07</th>
<th>Nov-07</th>
<th>Dec-07</th>
<th>Jan-08</th>
</tr>
</thead>
<tbody>
<tr>
<td>rdf</td>
<td>20.97%</td>
<td>11.11%</td>
<td>18.07%</td>
<td>28.00%</td>
<td>20.63%</td>
<td>17.86%</td>
</tr>
<tr>
<td>semanticweb</td>
<td>12.90%</td>
<td>5.56%</td>
<td>13.25%</td>
<td>20.00%</td>
<td>9.52%</td>
<td>8.93%</td>
</tr>
<tr>
<td>programming</td>
<td>4.84%</td>
<td>7.41%</td>
<td>6.02%</td>
<td>4.00%</td>
<td>4.76%</td>
<td>5.36%</td>
</tr>
<tr>
<td>redland</td>
<td>9.68%</td>
<td>5.56%</td>
<td>7.23%</td>
<td>12.00%</td>
<td>7.94%</td>
<td>5.36%</td>
</tr>
<tr>
<td>library</td>
<td>9.68%</td>
<td>11.11%</td>
<td>13.25%</td>
<td>8.00%</td>
<td>4.76%</td>
<td>7.14%</td>
</tr>
<tr>
<td>xml</td>
<td>4.84%</td>
<td>5.56%</td>
<td>4.82%</td>
<td>8.00%</td>
<td>1.59%</td>
<td>3.57%</td>
</tr>
<tr>
<td>python</td>
<td>3.23%</td>
<td>0.00%</td>
<td>3.61%</td>
<td>4.00%</td>
<td>4.76%</td>
<td>3.57%</td>
</tr>
<tr>
<td>web</td>
<td>1.61%</td>
<td>0.00%</td>
<td>1.20%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>1.79%</td>
</tr>
<tr>
<td>framework</td>
<td>1.61%</td>
<td>0.00%</td>
<td>2.41%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>1.79%</td>
</tr>
<tr>
<td>semantic</td>
<td>0.00%</td>
<td>1.85%</td>
<td>1.20%</td>
<td>8.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>software</td>
<td>1.61%</td>
<td>1.85%</td>
<td>1.20%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>3.57%</td>
</tr>
<tr>
<td>opensource</td>
<td>0.00%</td>
<td>1.85%</td>
<td>1.20%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>api</td>
<td>0.00%</td>
<td>1.85%</td>
<td>3.61%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>3.57%</td>
</tr>
<tr>
<td>development</td>
<td>1.61%</td>
<td>0.00%</td>
<td>1.20%</td>
<td>0.00%</td>
<td>1.59%</td>
<td>1.79%</td>
</tr>
</tbody>
</table>

None of the ten URLs identified in Table 1 can be said to demonstrate stability of tag use over time. Over the period of time represented by this dataset, all of the ten URLs evinced a pattern similar to that of found for http://www.w3c.org/RDF for the period January 2007 to January 2008. As shown in Table 3, the percentages of occurrence for each of the top seven tags varies from month. Given that this URL points to the W3C’s RDF homepage, it is to be expected that the tag rdf would dominate. However, with the exception of rdf, the tags assigned to this resource demonstrate variability of occurrence during this period. For example, the tag semanticweb accounts for anywhere from 4.48% to 15.63% of total tag assignments during this period.

Even though there is no overall indication of emerging stability of tags over time, there are some URLs that appear to be on a path toward stability. Table 4 lists the 14 tags that account for 68.3% of the total number of tags assigned to http://libRDF.org. In the five-month period from August through December, 2007, a growing stability can be seen among the top seven tags appear to demonstrate a growing stability as the remaining tags that had originally accounted for 68.3% of tag use appear to be phased out. However, in January 2008, this seeming path to stability is interrupted as tags that were not used regularly, if at all, during the previous five months are resurrected.

Convergence

As defined in this study, stability is an indicator that tags are being assigned in a systematic manner. As such, stability may point to the possibility of convergence of the tagging vocabulary. Convergence occurs when, over time, the total tagging vocabulary narrows to — or converges on — a core set of tags that are assigned to the resource. Like stability, convergence is operationalized as a proportion. However, the convergence ratio is the proportion of total occurrences during a given month of the tags accounting for 68.3% of tags to the total number of tags assigned during that month. Convergence occurs when the ratio ramps up toward 1.0. However, because of the sampling method used (i.e., since the tags supplying the top 68.3% will generally account for almost 70% of the total tags assigned), convergence can only be assumed when the convergence ratio is appreciably greater than .7.

Only three of the ten resources identified in Table 1 can be said to display the type of stability associated with the data presented in Table 4 (i.e., http://libRDF.org,
Figure 1. Convergence ratios for http://libRDF.org from September 2004 through February 2008

Figure 2. Convergence ratios for http://simile.mit.edu from June 2004 through February 2008

http://simile.mit.edu and http://www.foaf-project.org. Because these resources would seem to be most likely to demonstrate a stable vocabulary and thus offer the best chance for exhibiting convergence in the tagging vocabulary, they were selected for further analysis. However, due to space limitations, only http://libRDF.org and http://simile.mit.edu are discussed here.

Figure 1 shows the convergence ratios for http://libRDF.org from September 2004 through February 2008. There may seem to be evidence of convergence in the ratio peaks near or in excess of .9; however, these are simply illusions of convergence in that these high convergence ratios do not persist but are frequently followed by dramatic drops to less than .6. Figure 2 shows the convergence ratios for http://libRDF.org from June 2004 through February 2008. The convergence ratios for http://simile.mit.edu demonstrate less variability across time, but do not come near the levels of convergence that were seen with http://libRDF.org.
Conclusion
Unlike previous research, this study has adopted a longitudinal approach to the evaluation of tagging practices. While it seems that there is a sometime appearance of stability in the application of tagging vocabularies in del.icio.us, further evaluation shows that these localized events of stability are, in fact, transitory and fleeting and are not associated with convergence of the vocabulary on a core set of tags common to all taggers.

References
A Visualization Software Tool for Library of Congress Subject Headings

Abstract
The aim of this study is to develop a software tool, VisuaLCSH, for effective searching, browsing, and maintenance of LCSH. This tool enables visualizing subject headings and hierarchical structures implied and embedded in LCSH. A conceptual framework for converting the hierarchical structure of headings in LCSH to an explicit tree structure is proposed, described, and implemented. The highlights of VisuaLCSH are summarized below:
1) revealing multiple aspects of a heading; 2) normalizing the hierarchical relationships in LCSH; 3) showing multi-level hierarchies in LCSH sub-trees; 4) improving the navigational function of LCSH in retrieval; and 5) enabling the implementation of generic search, i.e., the ‘exploding’ feature, in searching LCSH.

1. Introduction
Library of Congress Subject Headings (LCSH) serves as a de facto universal controlled vocabulary for libraries in many countries (O’Neill & Chan, 2003). Beyond its use in traditional libraries, LCSH also serves as a tool for interoperability among metadata schemes (Vizine-Goetz, et al., 2004), for creating cataloguing records of digital information (Frank & Paynter, 2004; Riley, 2007), and for integrating multiple information depositories (Koch & Day, 1997). Despite these expanded applications of LCSH, most users underutilize LCSH’s full power in part due to its complexity and the lack of an effective tool to guide users.

The conversion of hierarchical term references in the entire LCSH (called ‘hidden classification’ of LCSH) into an explicit tree structure has long been desired (Richmond, 1959). To date, there is no existing tool for displaying the complete ‘hidden classification’ of LCSH. Two popular systems for searching LCSH and LC authority records (source of LCSH) are currently available: a) the LC Authorities website (http://authorities.loc.gov) which is freely open to the public; and b) the OCLC Connexion service (http://www.oclc.org/connexion) which is accessible only to registered users. Both systems are primarily designed for accessing a single authority record at a time. To view the hierarchical structure surrounding a heading, users are required to take additional steps to visit each heading individually. In the recently released report on the future of bibliographic control, recommendations for the optimal use of LCSH include: 1) “transform LCSH into a tool that provides a more flexible means to create and modify subject authority data”; 2) “[e]nable its customized assembly into topical thesauri”; and 3) “evaluate the ability of LCSH to support faceted browsing and discovery” (Library of Congress, 2008, 35).

The following functions, valuable not only to general users but also to professionals responsible for development and maintenance of LCSH including the display of ‘hidden classification’, are only partially supported by current LCSH tools: 1) displaying the full hierarchical structure of LCSH; 2) showing multiple locations of a LC subject heading in LCSH hierarchies; and 3) displaying LCSH terms all of which are synonym/homograph.

The aim of this study is to develop a software tool for effective searching, browsing, and maintenance of LCSH by enabling the visualization of subject headings and hierarchical structures implied and embedded in LCSH. This new LCSH visualization tool (called VisuaLCSH) is proposed to fill the aforementioned gaps in current systems. It is targeted
for various user groups, from novices to professionals, and is applicable to a wide range of applications including the searching, browsing, updating and maintaining of LCSH, as well as allowing for the integration of LCSH to other knowledge organizations.

A primary function of the VisuaLCSH is to enable the visualization of the full hierarchical structure of LCSH in both macro and micro views, thereby enabling effective browsing and traversing of LCSH. VisuaLCSH also offers smart searching of LCSH by considering both the synonyms and the word-based or character-based matches of search terms. The major features supported by the LCSH visualization tool include: 1) the visual outlook of the headings, concepts, and structures of LCSH; 2) the traversal of LCSH via zooming-in and out; 3) the smart searching capability; 4) the ability to locate multiple positions of specific subject headings across the entire LCSH hierarchical structure; and 5) the ability to demonstrate the synonymy and polysemy control of LCSH. The aforementioned features will be demonstrated with VisuaLCSH in a later section.

2. LCSH Tree
For this study, the hierarchical relationships of LC subject headings are presented in a tree structure, with the tree representing the entire LCSH being called LCSH Tree.

2.1 Conceptual Framework for LCSH Tree
A tree is a well-known data structure representing a hierarchical structure (Cormen et al., 2001). It consists of a set of nodes with a special node (called root node) placed at the highest level of the tree and links between the various nodes. There is only one root node per tree. A node placed one level immediately higher (lower) than the current node is called the parent (child) node. A node that does not have any child nodes is called a leaf node, and is the last node along the path starting from the root node. The root node of a tree is the tree’s only node that does not have a parent node.

Let LCSH Tree be a tree in which each term in LCSH is denoted by a tree node. Four types of terms can be found in a subject authority record: established terms, Related Terms (RTs), terms for Used For (UF), and Broader Terms (BTs). In LCSH, hierarchical relationships between terms are not always rigorously defined, particularly with regard to subject heading strings (i.e., combinations of main heading with one or more subdivisions). As a result, this study focuses on LC main headings only, due to difficulties involved in determining broader and narrower relationships among subject heading strings involving multiple facets.

A hierarchical structure can be built in such a way that a relationship between an established term and its BTs are denoted by child-parent nodes in the tree. RTs and terms for UF are excluded from constructing the hierarchical structure in the tree because they are not in a hierarchical relation with established terms. UF terms are included solely as access points in the searching of the VisuaLCSH system.

Once the algorithm for constructing a LCSH Tree is applied to the LC subject authority file, the hierarchical structure in the authority file is represented by a set of n trees. Each of the n trees is independent of the remaining n-1 trees, meaning that no hierarchical relation (link) exists across different trees. Each of the n trees is called a concept tree as it represents a concept. To connect all the n trees into one, an artificial node is created and used as a parent node for all the collective n trees. The resultant single tree becomes the LCSH Tree where the fictitious node becomes its root node. Hence, the LCSH Tree consists of n trees called LCSH sub-trees as they are all located under one root node. The
complex structure of the LCSH Tree (concatenation of n LCSH sub-trees) constitutes the hierarchical structure of LCSH.

2.2 Development of VisuaLCSH

In this study, a retrospective version of the LC subject authority file, consisting of 291,000 authority records and covering the years 1986 to 2005 available via the Library of Congress Cataloging Distribution Service, is utilized to build a LCSH Tree. Other resources, such as bibliographic records, are not considered because we want to build the Tree based upon subject headings authorized by LC only so that we have a set of subject headings validated by a single authority. The subject headings for topical terms (specifically headings in the MARC field 150 subfield code a) are used in building the LCSH Tree, excluding juvenile headings (beginning with sj).

An LCSH Tree is constructed based on the algorithm and dataset described above. The 2005 retrospective version of LC subject authority files produces 28,136 (equivalent to the n above) trees, i.e., 28,136 LCSH sub-trees. The LCSH Tree is composed of a concatenation of 28,136 trees and an artificial root node placed over the combined trees. The sizes of the LCSH sub-trees range from sub-trees consisting of one node to some with a few ten thousand nodes.

The LCSH visualization tool is developed to display the resulting LCSH Tree, equipped with text-based and structure-based searching features. The LC subject authority file is manipulated and imported into an Access database which becomes the experimental LC subject authority database. The visualization system for the LC subject authority file is implemented using Microsoft Visual Basic 6.0.

3. LCSH Visualization Tool: VisuaLCSH

Two key functions of the LCSH visualization tool are: 1) navigation of the LCSH tree; and 2) searching of LC subject headings. Some additional functions not fully implemented in pre-existing LCSH tools, which were discussed in the introductory section, will also be demonstrated in this section.

3.1 Browsing LCSH Tree

To browse the LCSH Tree, the LC subject authority Access database should be selected in the ‘Browsing LCSH’ section (see Figure 1). All the root nodes (top-level) of the 28,136 LCSH sub-trees will then automatically appear, in alphabetical order, under the ‘LCSH sub-trees’ section. Using scroll bars on the right and bottom, users are able to identify and locate top-level subject headings of the sub-trees. When a top-level subject heading is clicked, the corresponding LCSH sub-tree is displayed on the ‘Selected LCSH sub-tree’ section. Figure 1 shows a resultant screen when the ‘Child care’ top-level subject heading is selected. In the ‘Selected LCSH sub-tree’ section, the plus (+) symbol is shown in front of subject headings only when there exists at least one child node. By clicking the plus (+) symbol, the selected node will expand, leading to the display of the immediate child nodes of the selected node. On the contrary, when a minus (-) symbol is clicked, the sub-tree under the node with the minus (-) will be collapsed back into one node. By zooming-in and out using the symbols, every node in the selected LCSH sub-tree can be visited, and the full hierarchical structure of the selected LCSH sub-tree can be displayed.
3.2 Searching LCSH Tree

Figure 2 shows a VisualLCSH resultant screen when the term ‘child care’ is searched. There are two search options for matching: complete (or exact) word-matching and partial word-matching. With partial word-matching, matching between headings and search terms is made on a word-by-word (character-by-character) basis. For example, a search for ‘Child care workers’ will not match the term ‘Child care worker’ with the option of complete word-matching, but will retrieve it when using the partial word-matching option.

When a search term is typed in the ‘Searching LCSH’ section, the system identifies three LCSH sub-trees, each of which contains at least one node whose headings are com-
pletely word-matched with the search term. The three top-level subject headings (‘Child care’, ‘Child care workers’, and ‘Persons’) of the LCSH sub-trees are displayed in the ‘LCSH sub-trees’ section. Clicking on a top-level subject heading (‘Child care’ in Figure 2) triggers two subsequent events: 1) the hierarchical structure of the corresponding LCSH sub-tree appears in the ‘Selected LCSH sub-tree’ section; and 2) the complete paths from the top-level node to the nodes containing the search term are displayed in the ‘List of hierarchical paths to search term’ section. In the ‘Selected LCSH sub-tree’ section, the nodes matched with a search term are highlighted with a black background. In the ‘List of hierarchical paths to search term’ section, nodes of different levels are designated by the ‘>’ symbol with general headings on the left side of the symbol and specific headings on the right. The ‘I’ in parenthesis at the end of a path indicates that the attached node is an internal node (not a leaf node) of the sub-tree, whereas the ‘L’ designates a leaf node.

3.3 Multiple locations of a heading in LCSH hierarchies
Using the VisualLCSH search function, users are able to identify how many times a heading or search term appears in LCSH as well as where it is located within the hierarchy of the LCSH Tree. For example, results for the term ‘weblogs’ searched with the complete option show that ‘weblogs’ as a heading appears once in each of two LCSH sub-trees, of which the top nodes are ‘Auxiliary sciences of history’ and ‘Biographical sources.’ The paths to the nodes are: ‘Auxiliary sciences of history > Civilization > Social Sciences > Sociology > Communication > Information science > Information resources > Electronic information resources > Computer network resources > Web sites > Weblogs’ and ‘Biographical sources > Diaries > Weblogs.’

3.4 Multiple LCSH terms with synonym/homograph control
When searching normally, only main headings are used for matching. VisualLCSH, however, provides the option of searching synonyms via UF references. With the synonym option on, both the main heading and its UF references are taken into account in the search process. Let us repeat the same search conducted in Figure 2 with the synonym option on. This time, an additional heading (‘Nanny placement agencies’) is retrieved as a result of the search because a UF (‘Home child care workers placement agencies’) contains the search term (‘child care’) even though the heading itself does not contain it.

By using VisualLCSH, users can also benefit from the terminological control of homographs in LCSH. Type-in ‘mercury’ as a search term with default search options (complete word-matching and no synonym searching) to see how the homographs of the term are controlled. The search reveals the following results, each of which presents a different meaning of the word ‘mercury’:
– Minerals > Ores > Metals > Liquid metals > Mercury
– Transportation, Automotive > Automobiles > Ford automobile > Mercury automobile
– Gods, Roman > Mercury (Roman deity)

Note that other homographic examples, such as ‘Mercury (Planet)’, are not searchable with the current version of this system as the heading is defined under MARC field 551 (heading — geographic name) which is not included in the authority Access database.
3.5 LCSH’s irregular hierarchical structures
A well-recognized problem with LCSH is its irregular hierarchical structure: “the relationships [of equivalence, association, and hierarchy] are inconsistent and may not exist at all on older terms.” (Library of Congress, 2008, 34) This new visualization tool does not point them out automatically. It greatly helps, however, in identifying where such flaws occur, especially when there are suspicious locations or terms. For example, while inspecting the lineage of mammals, the conflicting hierarchical structures embedded in LCSH are found as follows:

– Life (Biology) > … Vertebrates > Mammals > Primates > Cercopithecidae
– Life (Biology) > … Vertebrates > Mammals > Primates > Monkeys
– Life (Biology) > … Vertebrates > Mammals > Primates > Monkeys > Cercopithecidae

According to the first and third chains of subjects, the subject Cercopithecidae is immediately connected to both the subject Primates and to the subject Monkeys on different levels of the hierarchy. If the first and second chains are true, then the third must be false. If the second and third are true, however, the first must be false. Thus, two existing contradictory lineages are identified.

Another example of irregularity can be found in a search on ‘literature.’ It can be expected to see a single LCSH sub-tree beginning with the ‘literature’ heading on the top. However, the system returns approximately 100 different LC sub-trees, each of which appears to represent the literature of a specific region or genre. For example, headings at the top-level in the sub-trees include: ‘Greek literature’; ‘Haitian literature’; ‘Romani literature’; and ‘West African literature (French)’. There are no hierarchical links between any of the subject-trees, thereby indicating that they are separate concepts in terms of hierarchy, which is not truly the case.

4. Conclusion
This project proposes a conceptual framework of LCSH and develops a new tool for visualizing the structure of LCSH. Major functions of the new tool are described and demonstrated above, including features that have not been embedded in existing tools. The highlights of VisuaLCSH are summarized below: 1) it reveals multiple aspects of a heading; 2) it is a tool for normalizing the hierarchical relationships in LCSH; 3) it shows multi-level hierarchies of terms in LCSH sub-trees; 4) it improves the navigational function of LCSH in retrieval; and 5) it enables the implementation of generic searching, i.e., the ‘exploding’ feature, in LCSH.

The conceptual framework shown in the LCSH Tree and VisuaLCSH may contribute to the utilization, development, maintenance, and application of the current LCSH in the following ways: 1) Enhancing subject searching in LCSH: In most cataloguing systems, one can search a heading or a word, but not the relationships between headings (Weinberg, 1998, 371); VisuaLCSH provides an enhanced means of searching by presenting the complete hierarchy of LCSH to users; 2) Building a hierarchy: The developed algorithm and software may be applied to other controlled vocabulary lists in which a hierarchy is embedded; the hierarchical display facilitates the generation of mini-thesauri based on a subset, such as LCSH sub-trees of the full structure; 3) Maintaining and controlling quality: This tool can be utilized to review and verify the hierarchical structure of LCSH; 4) Automating organization/classification based on LCSH: LCSH-based automated classification can play a central role in building digital libraries using traditional library classification schemes (Koch & Day, 1997; Yi, 2005); the LCSH concept trees can be used...
as a basis for organizing electronic resources; 5) Mapping to folksonomies: Facilitating mapping of LCSH to folksonomies or user-assigned tags; 6) Enabling interoperability with other controlled vocabularies: A number of previous studies attempted to link LCSH to other controlled vocabularies (Vizine-Goetz et al, 2004); the notion of concept trees may be useful in future interoperability research.

Acknowledgement
The authors would like to thank unknown reviewers for their valuable comments and suggestions.

References
Hypertext Model — HTXM
A Model for Hypertext Organization of Documents

Abstract
This article reports an applied research on the construction and implementation of a semantically structured conceptual prototype to help in the organization and representation of human knowledge in hypertextual systems, based on four references: the Facet Analysis Theory (FAT), the Conceptual Map Theory, semantic structure of hypertext links and the technical guidelines of the Associação Brasileira de Normas Técnicas (ABNT). This prototype, called Modelo Hipertextual para Organização de Documentos (MHTX) — Model For Hypertext Organization of Documents HTXM — is formed by a semantic structure called Conceptual Map (CM) and Expanded Summary (ES), the latter based on the summary of a selected doctoral thesis to which access points were designed. In the future, this prototype may be used to implement a digital library called BTDECI — UFMG (Biblioteca de Teses e Dissertações do Programa de Pós-Graduação da Escola de Ciência da Informação da UFMG — Library of Theses and Dissertations of the Graduate Program of School of Information Science of Universidade Federal de Minas Gerais).

Introduction
Since hypertext documents appeared in the last decade, specialized literature stresses that one of the main problems related to the use of this new tool is inefficient orientation of the user when consulting the document. This situation is often caused by the lack of clear and consistent criteria in the process of conceptual organization of the hypertext. This is especially true in the phases of choosing and structuring the relevant information and in the creation of links from the existing relationships between them.

The work of semantic organization is an essential requisite in the creation of hypertext systems. It allows that in a later phase either the user or the search tools are more efficient when information is to be accessed. A disorganized navigation path due to a conceptually deficient hypertext design tends to make search and browsing more difficult for the user, and may even jeopardize the retrieval through search engines.

Semantic organization of a certain area of the knowledge for creation of a hypertext system would have a semantic or conceptual net as a product; and this would require the implementation of a system of concepts integrated between themselves. Ideally, this process should precede the phase of implementation of the system, occurring in the phase of hyperdocument construction.

According to Schiper, cited by Lucas (1995, p. 70), hypertext is structured according to the logic of knowledge and in principles of psychological reasoning. However, a question remains about the dynamics of the meaning of the texts, in which different readings and associations on interpretation, comprehension, utilization and applicability of the information in the hypertext system can occur.

Thus, it can be assumed that if information is to be displayed in a logical and organized manner, it will probably be easier to locate and retrieve what one is looking for in a hypertext. The ideal scenario would be that the semantic organization of the text, which can be called conceptual structure, would reflect the semantic relationships between the subjects that are comprised in the texts.

The importance of conceptual modeling during construction of the hyperdocument is clear since it allows a more efficient retrieval. Nonetheless, given the enormous amount
of documents already available on the Web, their distribution and plurality of the products and their producers, a regular and precise construction of semantically structured hypertexts is utopian.

A significant segment of the literature points out to the relevance of the analytical-synthetic characteristics of the facet analytical theory (FAT) of Ranganathan as a possible solution in conceptual modeling of hypertext systems. Having in mind the possible creation of a conceptual modeling as a component of a prototype to help authors of hypodocuments and information professionals in the transformation from linear text to hypertext, the importance of studying the facet analytical theory is justified.

Theoretical and methodological referentials
In the process of creation of this model four referentials were considered:

1. The procedures of the Facet Analysis Theory (FAT), formulated by S. R. Ranganathan (1967), later studies of the Classification Research Group-CRG (1985) and Spiteri (1998); they make use of the analytical-synthetic methodology of knowledge organization, which allow the user to see subjects of a document under different points of view aiming at a knowledge map of a certain area, be it scientific or not. FAT, as a classification and indexing technique, supports the a priori knowledge organization, having semantic structure modeling as the starting point.

2. Concept mapping which is a proposal of Joseph D. Novak (2002) for the education area. The proposal, however, is based on the significant learning theory by education psychologist David Ausubel (1963, 1968, 1978). As visualization tools, concept mappings make the structuring of documents published as hypertexts easier since knowledge to be retrieved is made available in a user friendly manner. The information representation and retrieval potential of concept map is seen as a navigation alternative for the virtual text.

3. The concept of hypertext link, gradually developed by Bush (1945); Conklin (1987); Frei and Stieger (1995); Baron et al. (1996). It allows showing the relationships between the concepts and, as a navigation tool, also uncovers the existing types of semantic relationships (hierarchical, associative) between the concepts, giving more flexibility to the proposed model.


In the elaboration of the Hypertext Model — HTXM, the recommendations of Campos (2001) were followed. In the first level (Level of understanding of subject approach), the following requisites were defined: (1) Determination of the knowledge domain: Organization of knowledge; (2) Method of reasoning utilized for representation of knowledge units: Facet analysis; (3) Type of reader: Academic community, researchers and professionals of Information Science and correlated areas; (4) Document typology of the hyperdocument: Theses and dissertations. In the second level (Level of organization of know-
ledge units in the construction of the narrative) the following requisites were defined: (5) Nature of the content of knowledge units: Modularity; (6) Establishment of relationships between knowledge units: Creation of links. In the third level (Level of establishment of a communication and expression vehicle about the hyperdocument theme) the following requisite was defined: (7) Elaboration of a graphic representation for conceptual nodes and their relationships: Graphic implementation of the HTXM.

**Creation of the HTXM (Hypertext Model for Organization of Documents)**

The digital prototype was named Hypertext Model for Organization of Documents (HTXM) and it has two elements: a semantic map called Concept Map (CM), and an Expanded Summary (ES), a tool formed by the summary of the thesis to which points of access are aggregated. The hypertext model was installed in a database in a digital format. It comprised the thesis selected as the object of this research and later it should contain complete digitalized text of all other theses and dissertations from the Graduate Program on Information Science of the Universidade Federal de Minas Gerais, Brazil.

One of the first procedures of this study was the selection of a doctorate thesis as a working material. The choice of a microtext was due to limitation of the scope proposed for the model which is to work on a specific document universe, in a specific subject domain. In later studies expansion of the application to a group of documents of the same nature is planned.

A thesis was chosen as a sample for conversion of a linear text into non-linear text because of the following criteria: (1) theses and dissertations are linearly written; (2) the difficulty of the thematic author to work directly with the technological author at the same time; (3) the impossibility of the theses being concomitantly written in linear and non-linear form; (4) the possibility of application of this model at the BTDECI — UFMG (Library of Theses and Dissertations of the Graduate Program on Information Science of UFMG.)

As a wider application of the model hereby presented, each thesis to be inserted in the database will have a conceptual structure that will show its content and a formal structure that depicts its sections (summary). Although the thesis used had unique characteristics such as typology and theme, the conceptual modeling proposed can be tested in the construction of other hyperdocuments, from its own characteristics.

The combination of the techniques of the approaches described above has served as the basis for creation of an interactive tool with an organizational potential of semantic content in complete documents in hypertext databases with the possibility of efficient context retrieval. Below, the constructions of the conceptual and formal structures are described for elaboration of the HTXM.

**The conceptual structure of the HTXM**

The methodological procedures for conceptual implementation included, sequentially, the following steps: (1) Identification of basis document for work (the thesis); (2) Reading of the thesis; (3) Facet analysis of the subject of the thesis: selection of relevant terms and categories (facets); subfacets recognition; ordering facets, subfacets and foci for the concept map, and finally, organization of all terms and their relationships; (4) Creation of the concept map (CM) with its links and relationships; (5) Structuring the Expanded Summary (ES); (6) Creating summary links for the text.
The facet analysis

For facetation of the thesis chosen for this study, the technique of facet analysis was utilized, based on two processes: (a) analysis, which begins during identification of relevant concepts; (b) classification of concepts into categories, in which each category represents one characteristic. In the synthesis process each concept from the categories is combined with another to express the compound subject. It is worth noting that the synthesis was not implemented for the present thesis.

To form categories, the normative principles of the idea plan from the Simplified Model for analysis of facets: Ranghanathan 101, elaborated by Spiteri (1998) were utilized. In this work the author discusses and tries to synthesize previously established principles in two research frontlines: those initially presented by S. R. Ranganathan and those presented later by the Classification Research Group-CRG in London (1952).

The concept map (CM)

The objective of the concept map (CM) of the HTXM is to allow the user a general view of the semantic structure of the chosen text through its graphic representation. This would make context semantic navigation easier, through digitalized sections and subsections of the hypertext database. The conceptual structure is composed of relevant concepts of the chosen thesis, and organized according to the faceted structure with the semantic, hierarchical and associative relationships.

The hypertext structure of the CM was implemented with the Star Tree Studio Software. This software has several applications to interpret and make fragmented information available in a hyperbolic graphical format. The concept map has a three dimensional semantic structure where all relevant concepts of the thesis and its semantic relationships are represented. Navigation of CM is characterized by existing inter-relationships between the concepts.

As in FAT, elaboration of a concept map implies learning to group concepts according to their perceptive traces and also to previously established categories, involving selection, ordering and grouping. In the elaboration of a concept map similarities and difference relationships are made clear. The denominations “array” and “chain” are utilized in facet analysis to show the genus-species and whole-part hierarchical relationships. Navigation through a concept map allows the end user to navigate through semantically related concepts, allowing a more precise retrieval.

The creation of links

For structuring conceptual links the list of all concepts inserted in the concept map was considered, with their respective addresses, marked on the thesis text. As each concept may have only one electronic path, each concept had its link directed to the section or subsection of the thesis in which its content had the highest relevance.

As for function, a relational type link was utilized. As the name suggests, relational links exhibit the more relevant information non-linearly, helping the user to choose between decreasing or increasing its path in information searches.

As for directing the user, the links utilized were classified as active or inactive, and were color-coded.
Navigation in the CM
The graphical navigator is a schematic representation of the hypertext structure. It aims to provide the user with friendly ways to locate specific information. The navigational structure of the CM demonstrates to the user how its total content is semantically organized and how they connect internally, making moving from one node to the other easier. It also shows that the document can be graphically represented by nodes of information and the links between them.

CM navigation is characterized by concept browsing and searching mechanisms. Access to thesis text highlights the searched keyword in bold. From there, the user will be able to navigate the text, with the option of going to previous or later chapters, or yet, access the map again.

Hyperbolic visualization in the navigation of the CM exhibits the concepts of the focused links very close to one another, giving a better orientation and selectivity between concepts. In fact, the first window of the hyperbolic navigator is very similar to the aerial view from a tree, with a single and central trunk, dividing into smaller branches. During the operation of this map, the elements situated at the information nodes decrease and increase exponentially, causing distortion such as a fisheye, allowing it to encompass huge structures.

Formal structuring of the HTXM
In addition to the previously described conceptual structure, the HTXM also contains a formal structure that comprises the classical part of the document, the summary, which includes its organizational structure. Theoretically, this part interconnects pre-textual information with textual and post-textual elements. The summary thesis was expanded as an element of the proposed model. It included points of retrieval access not listed in the original summary, but that were considered relevant. The conception of the ES aimed to provide the reader of the hyperdocument a higher level of detail on the information contained in the document. It included information that did not appear in the sections of the summary, creating access links for these information and improving the proposed hypertext map.

The Expanded Summary (ES)
The structure of the summary is presented according to the technical standards of the Associação Brasileira de Normas Técnicas — ABNT (Brazilian Association of Technical Standards), the NBR 6024/2003 (formerly NBR 6027/1989) — Summary/Presentation. The analyzed thesis has 10 chapters divided into secondary and tertiary sections, also exhibiting a list of tables, Portuguese and English Abstract and four appendices: Appendix 1: Interviews; Appendix 2: Verbal Protocols; Appendix 3: Glossary; and Appendix 4: Copies of Sociology and Botany texts. For construction of the ES, some rules were defined: (1) Maintenance of the same format for titles utilized by the author in the summary topics; (2) Insertion of sections that were not included in the subdivisions of the summary in the ES and tables and figures in the context of their utilization and localization in the chapter to which they refer to.

The Greenstone Digital Library Software GSDL 2.50 (described below) was also utilized to create the sequential navigation structure of the ES of the thesis.

The formal structure of the HTXM corresponding to the ES was created from the summary of the thesis, with all its subdivisions, plus other information relevant elements for
the understanding of contexts not inserted in the concept map; this would compensate for their absence in the structure of the summary. At that point, organizational type links were utilized. This allowed to describe the superficial structure of the documents, comprising all the elements of the syntactic macro-structure of the summary presentation. For each main division and subdivisions URLs links were created, always inside each primary subdivision. To each added link that was not present in the primary, secondary or tertiary sections of the summary a sequential number was given; it corresponded to the section to which it belonged, in a corresponding numbering, to facilitate mnemonization of the summary structure. In this hierarchical structure, the information was, then, introduced from the general to the specific.

Technological implementation: constructing the prototype
After conceptual implantation of the prototype, development of this study reached a second phase: the technological implementation of the HTXM. Prototype is a visual representation of implementation of a project aiming to help the analysis of all functional needs (Quinn, 2002, p. 1). It is important because it allows, before testing, to generalize and disseminate the results, to see the constructed structure and to analyze the users’ needs. Due to its exploratory nature prototypes generally do not include the whole content of its application potential. Thus, the proposed prototype was constructed by using only the information that were necessary to test the HTXM. The components were the database where the thesis was stored, the concept map (CM) and the expanded summary (ES). The efficiency of the prototype was tested without the traditional delineation of an experimental research. Due to time and scope limitations, systematic experimental tests with users not involved in the research will only be done in later phases.

Optimization of technological tools was attempted during the process for selection of software for implementation of the HTXM. The Greenstone Digital Library software is specific for creating digital libraries and, despite some limitations, it allows the creation of complete texts databases. The Star Tree Studio software utilized in the construction and in hierarchical structuring of the CM also served for demonstrating hierarchical relationships but also associative relationships, allowing the insertion of more than one term in the same level.

Final remarks
Some problems found during the research are worth of note. The development of the prototype showed that there was a lack of technological support for structuring. The difficulties of choosing and coordinating between the two software could be minimized in future applications with possible different software packages that could implement all phases of the HTXM. It should be stressed that, in addition to the applications created in this thesis, the Greenstone software has characteristics that will allow to generate enough applications to set up a digital library, such as that of BTDECI — UFMG. On the other hand, the software has limitations for inserting applications in tools constructed with other software thus making it necessary to create links to redirect to those applications. This has created difficulties in the CM direct access. The CM had to be hosted inside the ES which forces the user to pass through an intermediary page before accessing it.

It can be concluded that the technique of facet analysis was efficient in the conceptual modeling of the thesis, providing methodical dynamics, from identification of relevant terms to formation of categories. CM provided an alternative for the problem of user
disorientation, although it did not include the technique of navigation path. On the other hand, a color-coded oriented navigation representing each hierarchical level of the map was created, showing to the user how all the semantic content is organized and how it is internally connected, facilitating hyperbolic navigation.

References
Information Retrieval from Digital Libraries
Assessing the Potential Utility of Thesauri in Supporting Users’ Search Behaviour in an Interdisciplinary Domain

Abstract
The objective of this research was to investigate the extent to which thesauri have the potential to support the search behaviour of nanoscience and technology researchers while interacting with an electronic book digital library. Transaction log data was obtained from a nanoscience and technology digital library to investigate the nature, type and characteristics of users’ queries and search terms. The specific objectives was to assess the extent to which users’ search terms matched with those found in two well-established thesauri attached o the INSPEC and Compendex databases.

Introduction
Nanoscale science and technology have seen rapid growth and expansion in new areas in recent years. Due to the interdisciplinary nature of this area, nano researchers need to consult a variety of resources. Kutner (2000) examines challenges and issues in searching in interdisciplinary areas. She argues that although electronic bibliographic databases that provide sophisticated searching capabilities and multiple access points to the scholarly literature have been a boon to the interdisciplinary researcher, problems continue to exist in terms of lack of consistent interfaces and consistent controlled vocabularies across databases. Although there are a number of studies focusing on search behaviour of interdisciplinary researchers, there is little research on whether thesauri have the potential to support interdisciplinary search behaviours. There is a growing interest in reusing and repackaging thesauri and other types of controlled vocabularies in various digital libraries, institutional repositories, content management systems and subject gateways. Controlled vocabularies such as the Art and Architecture thesaurus, ERIC, MeSH to name a few have found their way into web-based systems and services (Shiri and Revie, 2000; Shiri, 2006). In order to assess the utility and reusability of thesauri in new information environments, research needs to explore the type, nature and characteristics of user terms and queries within the context of new information repositories such as digital libraries. This type of research will shed light on the ways in which thesauri can be reused and/or redesigned to accommodate users’ searching needs. This study has investigated users’ queries and search terms as revealed by transaction log analysis of a nanoscience and technology digital library. Moreover, it has assessed the extent to which the search terms have exact or partial match equivalents in two thesauri attached to the INSPEC and Compendex databases.

Methodology
The theoretical framework upon which this study was carried out can be found in a number of user-thesaurus interaction studies (Shiri & Revie, 2004, 2005). The methodological framework provides the details of research questions, the dataset used in this study, thesauri used and data analysis methods.
**Research Questions**

The following research questions were formulated to address the objectives of this study.

1. What types of queries do NanoNetBase users formulate? Subject searches vs. known-item searches?
2. What are the characteristics of the queries in terms of the number of terms?
3. How many of the user terms did match those found in the INSPEC and Compendex Thesauri?
4. How many percent of search terms are exact matches, how many percent are partial matches?
5. What are the characteristics of the queries in terms of the use of acronyms in the query formulation process?


NANOnetBASE is an e-book digital library for nanotechnology and nanoscience researchers. The full-text database consists of 45 titles that can be accessed in a variety of ways. The library provides a variety of search tools, such as Boolean operators, truncation, wildcard, stemming, fuzzy searches, field searching, phonic search, synonym search, and variable term weighting.

**Transaction Log Dataset**

Transaction log analysis of web search engines, intranets, and websites can provide insight into understanding of the information searching process of online searchers (Jansen, 2006). The data used in this study consisted of transaction logs from the NANOnetBASE, a digital library subscribed and used by a Canadian university. Transaction logs from July 2004 to October 2006 were examined. In total, 1921 transactions were analyzed. The transaction logs contained the following information: date and time of activity, user identifier (in the form of IP address), activity detail (query), activity type: book viewed, advanced search, quick search, search terms, search results viewed, or managed account information.

**Thesauri Used**

Two thesauri were utilized for this study, namely INSPEC and Compendex. These thesauri are part of the INSPEC and Compendex databases. The reasons for the choice of these thesauri are a) they are both well-established thesauri and b) the cover many aspects of nanoscience and technology. The users’ search terms derived from transaction log analysis of NanoNetBase were all compared against these two thesauri.

**Data Analysis**

Three levels of analysis were defined, namely term-level analysis, query-level analysis and session-level analysis. For the purpose of this paper we will particularly focus on the first two levels. A list of search terms and queries was gleaned from the transaction log data set to allow for the analysis. All of the terms were compared with those in the INSPEC and Compendex thesauri.

**Results**

The results below are arranged based on the order of the research questions.
RESEARCH QUESTION 1: *What types of queries do NanoNetBase users formulate? Subject searches vs. known-item searches?*

<table>
<thead>
<tr>
<th>Query types</th>
<th>Number of queries</th>
<th>Percentage of queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject queries</td>
<td>333</td>
<td>84.74%</td>
</tr>
<tr>
<td>Title queries</td>
<td>32</td>
<td>8.14%</td>
</tr>
<tr>
<td>Author queries</td>
<td>3</td>
<td>0.76%</td>
</tr>
<tr>
<td>Unknown</td>
<td>25</td>
<td>6.36%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>393</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>

Table 1. Types of user queries

As Table 1 shows, subject queries constitute around 85% of the queries submitted by users. This is particularly important as it shows that subject queries constitute a major portion of user queries in the context of digital libraries.

RESEARCH QUESTION 2: *What are the characteristics of the queries in terms of the number of terms?*

Table 2 shows the number of queries along with the number of search terms used. The average number of terms used was 2.11 and the largest number of terms used was 8.

<table>
<thead>
<tr>
<th>Number of terms used in query</th>
<th>Number of queries</th>
<th>Percentage of queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>156</td>
<td>40%</td>
</tr>
<tr>
<td>2</td>
<td>137</td>
<td>35%</td>
</tr>
<tr>
<td>3</td>
<td>53</td>
<td>14%</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>5%</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>2%</td>
</tr>
<tr>
<td>6</td>
<td>13</td>
<td>3%</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>n/a</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Table 2. Number of query terms in searches

RESEARCH QUESTION 3: *How many of the user terms did match those found in the INSPEC and Compendex Thesauri?*

RESEARCH QUESTION 4: *How many percent of search terms are Exact matches, how many percent are partial matches?*

In order to assess the extent to which users’ terms mapped to the thesaurus, all terms entered by users were analyzed and their mapping situations are reported here. A brief description of different match types is provided below with a range of examples shown in Table 3.

**Exact matches:** While the notion of ‘exact match’ may appear to be self-explanatory, it should be noted that in our study singular noun terms which matched to the plural form in the thesauri were also treated as an exact match. Matching acronyms were also considered as being exact matches.

**Partial matches:** These types of match were typically between one query word and a multiword thesaurus descriptor. Partial matches also included user terms which had part-of-speech difference from thesaurus descriptors. This entails matched terms with different endings and terms that are adjectives of initial terms or have been subject to a stemming process.
Broader/narrower term match: these are the user terms that were mapped to a broader or narrower term in one of the thesauri used in this study.

Table 3. Examples of different types of match in INSPEC and COMPENDEX

<table>
<thead>
<tr>
<th>Match type</th>
<th>User term</th>
<th>INSPEC thesaurus term</th>
<th>COMPENDEX thesaurus term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact match</td>
<td>Vacuum</td>
<td>Vacuum (elementary particles)</td>
<td>Vacuum</td>
</tr>
<tr>
<td></td>
<td>Carbon nanotube</td>
<td>Carbon nanotubes</td>
<td>Carbon nanotubes</td>
</tr>
<tr>
<td></td>
<td>Sputtering</td>
<td>Sputtering</td>
<td>Sputtering</td>
</tr>
<tr>
<td>Partial match</td>
<td>Microchannel heat sink</td>
<td>Microchannel flow</td>
<td>Heat sinks microchannels</td>
</tr>
<tr>
<td>Narrower term match</td>
<td>Magnetic film</td>
<td>Magnetic thin film</td>
<td>Magnetic thin film</td>
</tr>
<tr>
<td></td>
<td>Tunnelling</td>
<td>—</td>
<td>Electron tunnelling</td>
</tr>
<tr>
<td></td>
<td>Photonics</td>
<td>Photonic switching systems</td>
<td>—</td>
</tr>
<tr>
<td>Broader term match</td>
<td>DNA scaffolding</td>
<td>DNA</td>
<td>DNA</td>
</tr>
<tr>
<td></td>
<td>vapour deposition stress</td>
<td>vapour deposition</td>
<td>vapour deposition</td>
</tr>
</tbody>
</table>

Table 4 shows the details of the user terms matched the INSPEC thesaurus along with the number of terms for each match type.

Table 4. Number and percentage of search terms entered by the user and matched in various ways to the INSPEC thesaurus

<table>
<thead>
<tr>
<th>Match type</th>
<th>Number of terms</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact match</td>
<td>152</td>
<td>37</td>
</tr>
<tr>
<td>Partial match</td>
<td>51</td>
<td>12</td>
</tr>
<tr>
<td>No match</td>
<td>174</td>
<td>43</td>
</tr>
<tr>
<td>Broader term match</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Narrower term match</td>
<td>23</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>408</td>
<td>100</td>
</tr>
</tbody>
</table>

The analysis in Table 4 shows that around 50% of the terms entered by users are matched those of the INSPEC thesaurus.

Table 5. Number and percentage of search terms entered by the user and matched in various ways to the Compendex thesaurus

<table>
<thead>
<tr>
<th>Match type</th>
<th>Number of terms</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact match</td>
<td>199</td>
<td>49</td>
</tr>
<tr>
<td>Partial match</td>
<td>53</td>
<td>13</td>
</tr>
<tr>
<td>No match</td>
<td>135</td>
<td>33</td>
</tr>
<tr>
<td>Broader term match</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Narrower term match</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>406</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 5 shows that around 62% of users’ search terms were matched those of the Compendex thesaurus.

A comparison of Tables 4 and 5 suggests that the Compendex thesaurus had exact matches for 49% of users’ search terms compared to 37% exact matches offered by INSPEC. This is particularly interesting as details and narrower terms of exact match terms can be further explored by users to find out more about the knowledge map of that particular subject area and can also choose a wider variety of terms offered by these thesauri.
A closer look at all the match types indicates that in general the Compendex thesaurus provides a better coverage for nanoscience and technology user terms. This is not to say that the Compendex thesaurus should be used instead of the INSPEC thesaurus. In fact, our observations of the nature of terms suggested by both thesauri show that there are terms whose exact or partial matches can only be found in INSPEC and not in Compendex. These observations suggest that both thesauri should be used in order to support users’ query formulation and expansion behaviours.

Acronyms

**Research Question 5:** What are the characteristics of the queries in terms of the use of acronyms in the query formulation process?

Acronyms were used in many of the queries. Around 10% (40) of the queries were acronym searches. This suggests that like many scientific fields, acronyms are well-used and intrinsic form of the language used within and to describe literature on nanotechnology.

<table>
<thead>
<tr>
<th>Acronym queries</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of queries using acronyms</td>
<td>9.67</td>
</tr>
<tr>
<td>Percentage of queries using 1 acronym by itself</td>
<td>4.33</td>
</tr>
<tr>
<td>Percentage of queries using acronyms and text words</td>
<td>4.58</td>
</tr>
<tr>
<td>Percentage of queries using 2 acronyms</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Since acronyms did not play a factor in the reformulated searches, it appears that the NANOnetBASE handles acronyms well and searches using them are able to retrieve results. However, since the database uses free text searching it is necessary that authors also use the acronyms in their writing. To ensure that all relevant documents are retrieved, it would be best to combine the acronym with its full-text equivalent using the Boolean operator OR.

**Conclusion and Discussion**

These findings suggest that these two thesauri have the potential to assist users in their search process, particularly for query formulation and expansion purposes. The results of this study suggest that Compendex and INSPEC thesauri can be reused in such a digital library as NanoNetBase to offer alternative search terms to users. From an information retrieval perspective, the INSPEC and Compendex thesauri have the potential to support both interactive and automatic query formulation and expansion processes. In the interactive query formulation or expansion process users can be offered a list of terms based on their initial search terms matched to one or both of the thesauri and then users will have the opportunity to choose from those terms or use broader, narrower or related terms. From an automatic query expansion perspective, users search terms can be automatically mapped to the terms and possibly include narrower terms of the ones with exact match cases. Another interesting observation in this study was that acronyms constituted around 10% of the queries submitted by users. This also suggests that there is a clear need to include acronyms along with full forms of terms in the thesaurus supported by cross referencing links to enhance the response rate of thesauri in supporting users’ queries. This research makes a number of contributions to the areas of knowledge organization, digital libraries and interdisciplinary search behaviour studies. The transaction data provides a very useful pool of terms using which a thesaurus can be built or the studied the-
Thesauri can be enhanced in the areas where there have been subject gaps. It will offer some implications for the design of thesauri and controlled vocabularies for interdisciplinary research areas. In particular, it identifies the areas that these thesauri lack in terms of subject coverage and specificity. Another contribution of this research is that it demonstrates the coverage of the INSPEC and Compendex databases and their associated thesauri in relation to nanoscience and technology. Since nanotechnology is a relatively new and multi-disciplinary field, this study will provide insight into interdisciplinary search behaviour and how users search and how the information access and retrieval interfaces may better be constructed so that users can access the information they need in an efficient and effective way.

Future Work
The findings from this study showed that thesauri such as the ones used in this study have the potential to support users’ search and interaction behaviour. Future research will focus on query reformulation and expansion cases within the dataset to explore the nature and types of query reformulation behaviour and the ways in which query reformulation and expansion will benefit from thesauri. Further analysis will be carried out to investigate whether we can identify any domain-specific search behaviours, such as the use of acronyms and initials or multiword search terms.

Acknowledgement
The authors would like to gratefully thank Randy Reichardt, Research Services Librarian (Engineering) University of Alberta, who provided us with transaction log data for the NanoNetBase e-book library.

References
Verónica Vargas — Mexican Institute of Water Technology
Catalina Naumis — Universidad Nacional Autónoma de México

Water-related Language Analysis
The Need for a Thesaurus of Mexican Terminology

Abstract
A need for uniformity in the Spanish terminology employed in Mexico on the subject of water per se and water issues in general and the lack of a controlled vocabulary are the main reasons behind this investigation, which seeks to establish the need for a terminological tool and to determine the feasibility of adapting an existing thesaurus. This research is based on a descriptive analysis of six thesauri that are in some way or another related to the subject of water. Emphasis must be made on the fact that water resource management is not the responsibility of only one area of science; it is multi and interdisciplinary. Furthermore, the heterogeneous nature of water resource management gives rise to considerable amounts of information that makes it necessary to study the area of knowledge concerned with the phenomena of information transmission and retrieval.

Introduction
Imagine you are asked to estimate the variation in time and space of water supply and demand at regional levels based on changes in land use, population growth and water industrialization. Or perhaps you are required to identify different possibilities for incorporating water valuation into an efficient management of water resources. These examples illustrate just two of the many issues faced by water managers and researchers as part of their daily responsibilities. The decisions that they make may affect numerous organizations, individuals, and the environment. Such decisions are based on a broad range of information gathered from a variety of reports that offer detailed scientific findings and evaluations. Many of these documents are authored by multi-disciplinary teams of experts who routinely use a wide range of terminology.

Experience has shown us that without a holistically-approached and controlled vocabulary on the subject of water in the Spanish used in Mexico, the indexer and the user do not have a reliable tool that will permit greater precision and ensure the necessary quality in indexing and information retrieval processes. In this regard, the inconsistency in the terms employed in the literature has lead to scattered information that represents an interesting information retrieval challenge. On the other hand, thesauri have proven to be useful tools that allow the possibility of indexing each of these heterogeneous information resources in a uniform way, enabling effective information retrieval.

As advised by most researchers while discussing the opportunity of constructing a thesaurus (ANSI/NISO 2003; Jorna, Davies 2001; Aitchison, Gilchrist, Bawden 2000; Currás 1998), a cost-efficiency analysis would confirm that, indeed, a new thesaurus is required. Moreover, because a thorough duplication of subject matter coverage is not common, “access to one or more thesauri in related fields can frequently serve as a useful starting point”. Even so, before embarking on the construction of a new thesaurus, it is essential to conduct an assessment of existing resources to ensure that the work is actually necessary.

It is important to take into account that different thesauri are available for different domains. Furthermore, even within a single domain, a variety of specialized thesauri are

employed that deal with specific areas. Because each particular field has its own terminol-
ogy, distinct thesauri are required. On the other hand, in order to exchange information
among different fields, a common thesaurus, which would likely be more general, would
be extremely advantageous.

Taking into account what we have mentioned up to this point, this paper presents the
methodology and results of the analysis of six thesauri related to water resources in order
to establish if an existing thesaurus could be adapted and used to index and retrieve holis-
tically approached water-related information using the terminology employed in Mexico,
or if it is necessary to create a new one, given that in Mexico such a thesaurus does not
exist.

Methodology and Results
As mentioned previously, water resource management is essentially interdisciplinary. An
integrated water resource perspective ensures that social, economic, environmental and
technical dimensions are taken into account in the management and development of wa-
ter resources. Keeping this in mind we have chosen six thesauri to include in a descrip-
tive analysis that may help us to establish a starting point on which to adapt an existing
thesaurus or build a new one. It is important to mention that a qualitative approach was
used in this study to analyze the context in which the term “water” was treated. Even if
our aim was to analyze each thesaurus, a comparison among them was necessary to de-
fine the main characteristics that a thesaurus for Mexico should have in accordance with
the informational necessities of a indexing system that contemplates water resources in
a holistic way.

In choosing the thesauri we held onto the fact that the main purpose of a thesaurus in
the retrieval of information is to provide a uniform and consistent vocabulary for index-
ing documents in information systems and to supply users with a specific vocabulary for
the retrieval of documents in such systems. A literature review indicates that controlled
vocabulary is crucial for the effective search and retrieval of information and that it also
greatly enhances the usability of any information retrieval system.

The thesauri chosen are:
1. Agrovoc (FAO)
2. General Multilingual Environmental Thesaurus (European Environment Agency)
3. OECD Macrothesaurus
4. Tesauro de Ingeniería Hidráulica (Hispagua)
5. Tesauro de Ingeniería Sanitaria y Ambiental (WHO)

The criteria for selecting these thesauri were the inclusion of the water resources topic,
seen from any and/or all perspectives: institutional backing of the thesaurus, taking into
account the experience and prestige of the thesaurus within scientific and academic fields;
and lastly, its availability and accessibility, a sine qua non requirement to be able to access
the thesaurus and carry out the corresponding analysis. We must mention that the first five
were available on line and that the sixth was only available in printed form.

2. The Water Resources Thesaurus, published by the Office of Water Research and Technology, U.S. De-
partment of the Interior, was not included in our research as it has been out of print for some time (1980). The
InterWATER Thesaurus, compiled by the IRC International Water and Sanitation Centre and partners, was not
available during our research.
The ANSI/NISO Guidelines for the Construction, Format, and Management of Monolingual Thesauri (2003) specifies four main purposes served by a thesaurus: 1) translation; 2) consistency; 3) indication of relationships; and 4) retrieval. As Owens and Cochrane (2004) mention, by evaluating a thesaurus we are able to find out if it meets these objectives.

In the absence of a tested and approved systematic methodology for the assessment of this kind of tools, we took into account the proposal made by Naumis (2002), which considers the following three essential elements in analyzing a thesaurus:

1. **Content analysis**: covers characteristics such as thematic fields, hierarchical list, semantic structure, thesaural relationships, reciprocal entries, preferred and non-preferred terms, scope notes, singular and plural forms, monolingual or multilingual, and maintenance.

2. **Distinctive characteristics**: backing of a responsible institution and its authority and experience in academic and scientific communities, the field of specialization, year published, up-dates, copyrights, introduction, and thesaurus structure (alphabetical display, subject display, and permuted index).

3. **Ease of use**: refers to the type of format (digital print), availability, accessibility, and usability.

In sum, we measured whether the subject coverage of the concepts displayed allowed for proper indexing and searching; if the structural relationships between the terms were adequately treated; and whether the thesaurus was easy to see, understand, and follow through on to proceed with efficient and effective indexing and searching.

While the analysis of the entire lexical content of the various tools that were made available to us would have been ideal, such an undertaking would have been impossible considering the resources available and the large number of terms contained in the thesauri we studied. The rather summarized observations that follow are based on a sample of all of the terms beginning with the word “water” in Spanish and in English in the monolingual or multilingual thesauri and in French in the Thésaurus Eau.

Based on the features that we mentioned of the thesauri analyzed, we are able to arrive at an initial inference that tells us it is unfeasible to adapt a particular thesaurus. However, it is possible to take the most relevant parts of these thesauri and adapt them to the Mexican reality. We found that of the two monolingual Spanish thesauri, one used a more universal vocabulary while the other was primarily local. The three multilingual thesauri employ a universal Spanish, although in some cases the translations are very literal and may cause more confusion in the terminology.

We noticed that terms are basically pondered from a technical standpoint and that the social and economic terminology required for a more interdisciplinary view of water resources management is missing. Most of the themes has to do with hydrology, hydraulics, and water supply, pollution, conservation, quality, chemistry, treatment, use, distribution, and so on. For example, none of these tools included the terms “water conflict” or “water rights”.

We must not forget that current water management has given rise to new perspectives on hydrological rationale. On the one hand, the sustainability paradigm has contributed new terms such as “integrated water resources management”, “water rights”, “water conflicts”, etc. On the other, globalization has generated a model of development that fosters the commercialization of water as a resource and the privatization of the water supply and
sanitation services, giving rise to terms such “water privatization,” “water economics,” “water stakeholders,” just to mention a very few.

We should not forget that some terms are just buzzwords and that they differ from jargon in that they have the function of highlighting or obscuring meaning, while jargon, ideally, has a well-defined technical meaning, if only for specialists.

The six thesauri we examined all share the usual forms of systematic display and alphabetical lists of descriptors. Five of them present a more specialized coverage, while only one can be described as truly general. The terminology can be described as formal and specialized for all six. In two of the thesauri we observed that proper names, even trademarks and names of government programmes, are included.

Most of the thesauri have an explicit relational structure connecting descriptors by relations of equivalence, hierarchy and various kinds of associations. However, the notation employed for the basic thesaural relationships are quite different in two of these tools. In a third one, the notation explained in the introduction differed entirely from the one that was used in the alphabetical display.

The effectiveness of tools for vocabulary management can only be maintained if the lexical and relational content is kept up-to-date. Updating a thesaurus is largely related to creating new descriptors. We observed that four thesauri mentioned in their introduction the importance of updating and invited users to suggest new terms. A fifth mentioned briefly its formal procedure. As López-Huertas (1998, 140) observes, “users’ search query terms should be considered as a source for the thesaurus vocabulary in addition to the traditional gathering methods of indexing documents, specialised vocabularies, other indexing languages and experts’ contributions.”

In other words, thesauri must be updated to incorporate terminology derived from the development of science or other fields, to fill the vacancies or lacks detected in them when they are in use, and to adapt them to the needs of the users as established by users’ searches.

Likewise, we noted that half of the thesauri were missing an introduction, which certainly makes an analysis more difficult. The lack of an introduction restricts the use of the thesaurus because, as Foskett mentions (1997, 133), an introduction briefly allows us “to cover all the sections: the limits of the field or fields covered, the choice and forms of terms, the network of relations, the form of presentation, and the use of special features.” Experience has shown us that even if the thesaurus can take various different forms, in every case it requires an introductory section to explain how best to use it.

Half of the thesauri analyzed are monolingual (two are in Spanish and one is in French), the rest are multilingual (three or more languages). We found that most of the multilingual translations from Spanish to English and vice-versa were fine. However, we must take into account that even if the handling of descriptors in multilingual thesauri is much the same as it is in a monolingual tool (Aitchison, Gilchrist, Bawden 2000), the use of foreignisms or deficient translations can cause confusion in the use of terms, which no doubt undermines the quality of the thesaurus and weakens its credibility. We also noticed that, unfortunately, an equal treatment of the languages is hard to achieve. As Jorna and Davies (2001, 286) observe “the complexity of natural languages, which all terminology is part of, rarely allows for one-to-one equivalences between terms of different languages.” Furthermore, technology cannot overcome the intellectual problem of semantic equivalence.
As we mentioned before, five of the thesauri studied were available online, attesting to the consolidation and expansion of the digital environment and confirming the need for greater availability of and accessibility to online thesauri. Consequently, it is no longer feasible to consider a printed version of a thesaurus exclusively. However, we found that due to the lack of standardization for publishing thesauri on the Web, a diversity of formats, structures and features are currently in use. This causes problems concerning thesauri interoperability, reusability and shareability. Based on our analysis, we found that there is an urgent need to examine semantic and syntactic tools, formats, and standards used by Web-based thesaurus publishers in order to find ways in which these aspects can be harmonized or integrated.

Conclusion

Based on the features of the analyzed thesauri that we have mentioned, we find that it is not possible to adapt an existing thesaurus and that a new thesaurus must be constructed that includes the terminology employed in Mexico and a more holistic approach to the water subject overall. Nonetheless, the structure of some thesauri may well serve as a foundation.

The varying perspectives through which the subject of water can be viewed must be recognized: from the fragmented compartments of scientific research to a more global and integrated social focus, water issues in Mexico and throughout the world demand a multidisciplinary, integrated and dynamic attention that encompasses both the social and scientific aspects of water resources as shown by the literature made available in recent years.

As we mentioned several times, thesauri are useful tools that permit the possibility of indexing all of the heterogeneous information resources in a uniform way, enabling effective information retrieval. The thesaurus to be developed must allow room for the terminology that users apply while searching for information and innovative means for documentary description must be explored through systematic indexation and through a more flexible information recovery adapted to the user. We cannot forget that access to information is much more generalized today and so we insist that much more user-friendly search interfaces be designed for the users. We must also not forget that the web display of the thesaurus is largely dependent on user requirements and that accessibility is also an important factor to note.

Thesauri should be revised regularly by means of well organised and standardised procedures. The inclusion of terms should be strictly controlled and documented to avoid any potential conflict with existing terms or structures.

We believe that modern thesauri should be multilingual in order to facilitate cross-cultural communication in an increasingly global information society. Furthermore, a multilingual thesaurus supports access to information resources not published in one’s native language.

Lastly, we must not forget that by fostering the use of information in society we can encourage the availability of and easy access to more information. The spirit behind information availability is to offer users everything they may require to satisfy their needs, independently of where they may be located. Accordingly, it is inconceivable to offer only a printed version of a thesaurus.
Acknowledgement
We would like to thank the Mexican Institute of Water Technology, in particular Ms. Patricia Navarro, for all of the support provided during this research.

References
Amanda Hill — University of Manchester

What’s in a Name?
Prototyping a Name Authority Service for UK Repositories

Abstract
This paper looks at approaches to name authority control in repository contexts and describes the work of the Names project, which has been funded to investigate issues surrounding the identification of individuals and institutions within repositories of research outputs in the United Kingdom.

Introduction
The problem of uniquely identifying authors has been with us ever since books have been catalogued. National libraries have been creating name authority files for authors of books for many years, starting with card catalogues and now maintaining electronic files in MARC format. However, authority files for the creators of journal articles do not tend to exist in library systems. The increasing use of subject-based and institutional repositories to hold working papers, reports, research data, and pre-refereed and post-refereed versions of articles has led to a corresponding rise in the number of authors identified in such systems.

Without having a means of uniquely and unambiguously identifying the creators of materials in repositories, it becomes difficult to be sure whether all the materials related to a particular author will be correctly associated with that individual. Names of authors may be entered in more than one way, or more than one author may have exactly the same name. This article looks at recent attempts to address this problem in the repository environment and goes on to explain the approach that is planned to be taken in the Names project.

Background
The context for the Names project is the work on repositories that has been undertaken in the United Kingdom in recent years, particularly in the rapid development of institutional repositories. One definition for these repositories is:

... a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members. (Lynch 2003)

Much of the development of these repositories in the UK has been seeded by funding from the Joint Information Systems Committee (JISC) through a series of initiatives that began with the eLIB programme in 1994 (Pinfield, 2004). As a result of the development of the ePrints software at Southampton which had been made possible by JISC funding, there were, by March 2003, six institutional repositories in the United Kingdom (Day, 2003), containing between them 7,334 records (it should be noted that 7,158 of these were in one repository: Southampton’s Department of Electronics and Computer Science, the first such repository to be established in the UK). Such has been the momentum behind the repository movement since then, that in early 2008 the OpenDOAR registry recorded 87 institutional repositories in the United Kingdom, holding 303,052 records.1 The number of records held within these repositories ranges from 179,722 at the University of Cambridge to the 26 repositories that are currently holding under 100 records.

1. The OpenDOAR directory of open access repositories is to be found at: http://www.opendoar.org/.
Subject-based repositories have been in existence for a longer period than institutional repositories. The longest-running is arXiv.org (covering e-prints in physics, mathematics, computer science, quantitative biology and statistics), which was first made available online in August 1991 (Ginsparg, 1994) and which now holds over 450,000 e-prints. RePEc is a service that covers working papers, journal articles and software relating to economics and this collection now holds over 560,000 items. RePEc itself benefited from JISC funding as part of the eLib programme in the late 1990s (Krichel, 1997).

Name authority control in repositories and related services
As might be expected, issues related to the reliable identification of authors were identified in the subject-based repositories once they reached a significant size. Searching by authors’ names has been among the top search methods by repository users. When a repository grows to substantial size, it is often the case that name variants cause headaches for both the users and repository managers. (Xia, 2006)

Where repositories contain relatively few items, the problems associated with loss of precision (the ability to retrieve only the items created by a particular individual) and loss of recall (the ability to retrieve all the items for that individual) may not be particularly noticeable and may be managed by the intervention of repository administrators. When repositories are large (or when the contents of different repositories are aggregated), these issues become more prominent. If author names are not controlled, then a search for a particular name will only retrieve items which match the query’s form of the name exactly, creating a loss of recall. If more than one author has the same name, then precision will also be affected, with irrelevant material being returned for a search.

Figure 1. illustrates this problem. A ‘Browse Authors’ search has been performed on the University of Cambridge’s institutional repository and the author names beginning with ‘A’ have been returned. It will be seen that the list reflects the differing ways in which authors’ names have been entered into the system. Some have been inverted, others have not, meaning that sometimes the ‘A’ is a forename, sometimes a surname. The name A.K.Bhushan appears in three different forms in the list and each of the names retrieves a different set of documents, though apparently all by the same person.

In order to avoid problems of this kind, the RePEc economics repository has introduced a level of authority control for its authors, in order to ensure that individuals are uniquely identified and correctly associated with the materials that they have made available through the services. Authors register with the service (this is also true for arXiv.org) before they can deposit materials and, as a result, RePEc currently holds information on 15,000 authors. The service also maintains a directory of economics departments and related institutions, which is linked to the RePEc Author Service to provide affiliation information for individuals.

This is how the service was described, in its early stages:

We take a set of digital library data, and we ask authors to tell us which papers they have written. Naturally, this strategy to get the authors involved in producing their own access control data will only work if the authors have incentives to supply such data. Since academic authors are interested in the visibility of their work, they will have good incentives to supply data to a database that is frequently consulted by potential readers. For any database to achieve such a status, it must be relatively large and available at low cost. The RePEc dataset of Economics research is a good candidate. (Cruz et al., 2000)

Variations on the author-registration approach pioneered by the RePEc service have since been adopted by commercially-run systems such as those supplied by Elsevier (Scopus Author Identifier\(^3\)) and Thomson Scientific (ResearcherID\(^4\)). ProQuest provides a more hand-crafted approach with its Scholar Universe service, which compiles lists of faculty members and associates them with their published works.\(^5\) Faculty are encouraged to add their own details to their profiles. Common to all of the systems is access to a large corpus of existing information about authors and their outputs which can be analysed and used to create associations between the entities concerned. They all also encourage authors to register and update their own lists of publications.

**The Names project**
With funding provided by the current JISC Repositories and Preservation Programme,\(^6\) the British Library and Mimas, a data centre at the University of Manchester, have been tasked with investigating:

---


6. Information on the programme can be found at: http://www.jisc.ac.uk/whatwedo/programmes/programme_rep_pres.aspx.
...the potential for the development of a Name Authority Service and factual authority for digital repositories, to support cataloguing, metadata creation and resource discovery in the repository environment. (JISC, 2007)

The British Library has a long history of expertise in the name authority area and is a contributor to the Name Authorities Co-operative (NACO). The Library of Congress (LC) maintains the LC/NACO authority file, which now contains around seven million records, created by over 500 contributors. The British Library is one of four permanent members of the Policy Committee of the Program for Cooperative Cataloguing which steers the governance of NACO.

Traditional library sources of authority control have proved to be inadequate for the purposes of institutional repositories. Even within the bibliographic world, Calhoun (1996) estimates that around 50% of author names in library catalogues are not represented in the LC/NACO authority file (LCNAF). In repositories the situation is more pronounced. The ePrints UK project and staff at MIT have both investigated the correlation between the Library of Congress name authority file (using services provided by OCLC) and author names held within institutional repositories. The results have not been published, but anecdotal evidence from those involved suggests that fewer than 25% of the authors depositing materials in repositories are also to be found in the LCNAF. It is clear that simply incorporating access to the LCNAF into repository software will not be sufficient for the purpose of uniquely and unambiguously identifying authors of repository materials.

Proposed approach

Initial work on the Names project has been investigating existing activities in this area and gathering requirements from stakeholders of the project. The stakeholders include repository managers, the project’s funders and partners, developers of repository software and providers of cross-repository services such as the Intute Repository Search service. These requirements are being used to develop the specification for the prototype that will be produced in the next phase of the project.

The prototype will be using a unique number-based identifier, rather than a controlled form of the name of an author (or an institution) as the primary identity key. It will also need to associate all known variants of the name with that number. In other words, the approach needs to be one of access control, rather than authority control. Such an approach might once have been seen by librarians as “brazenly radical” (Barnhart, 1996), but it makes perfect sense in a context where it is important to record the form of a name as it appears in a published article (to assist retrieval), even though this form may differ across different publications. As yet there is no agreed standard for the form of an author identifier and it appears unlikely that a single world-wide system for assigning a single identifier to an author will be implemented (Tillett, 2007). Therefore the prototype will need

---


8. The issue with name authority files is touched upon by MacKenzie Smith, Associate Director of Technology at MIT, in a mailing list message at: http://mailman.mit.edu/pipermail/dspace-general/2006-March/000902.html.

9. OCLC’s involvement with ePrints UK is described on the project page at: http://www.oclc.org/research/projects/mswitch.epuk.htm.

10. The Intute Repository Search service can be found at http://irs.ukoln.ac.uk/.
to be able to hold information about other identifiers that may be assigned to the same individual in related systems.

There are various sources of data that are available to the project team which will be used to populate the prototype. These include the author names and journal article information in the British Library’s Zetoc service\(^\text{11}\), which is an electronic table of contents covering journals and conference proceedings that have been published since 1993, and similar data from UK PubMed Central\(^\text{12}\), a repository for articles published in the life sciences (also maintained by the British Library) and from the metadata supplied to the repositories covered by the Intute Repository Search service mentioned earlier.

One of the activities undertaken by the British Library members of the project has been to analyse a variety of metadata schemas and standards and map the data types to the entities described in the International Federation of Library Association’s *Function-\al Requirements for Authority Data: a Conceptual Model* (FRAD).\(^\text{13}\) The schemas that have been mapped include the MARC21 format for authority data\(^\text{14}\), the National Library of Medicine (NLM) Journal Publishing DTD\(^\text{15}\), the Scholarly Works Application Profile (SWAP)\(^\text{16}\) and the International Standard for Archival Authority Records, Corporate, Personal and Family names (ISAAR (CPF))\(^\text{17}\). This will make it possible for a variety of different data sources to be adapted for use in the prototype and will ensure that the system will be able to export data in a variety of formats for use by other systems.

The requirements of repository managers are simple: they would like a name authority module that will plug into their existing repository software and provide auto-completion of author names for depositors of materials and for searchers of the system. It is possible for each repository to set up its own authority file for local authors, and functionality of this type is already offered by the latest version of the Eprints software.\(^\text{18}\) However, given the level of co-authorship across institutions, a centralised authority file for the UK would avoid duplication of effort. An international file would, of course, be even more useful, but that is beyond the scope of the current project.

Ensuring interoperability with repository systems and other services will be an important part of the development of the Names prototype and any future service that may evolve from it. This will involve the provision of a web service to enable support for auto-completion and may also entail allowing bulk harvesting of the Names data to trusted third parties. Enabling such functionality will also contribute to the wider, world-wide, efforts to unambiguously identify authors and institutions. The prototype’s web service will be tested during the course of the project by the Intute Repository Search and UK PubMed Central services to ensure that it meets their requirements for such a service.

The existing name authority services that were examined in the first part of this paper have all recognised that harnessing the knowledge of the authors themselves is an important part of providing a reliable name authority service. Authors are able to tell us

\(^{11}\) Zetoc is hosted and supported by Mimas and is available at http://zetoc.mimas.ac.uk/.
\(^{12}\) Available at http://ukpmc.ac.uk/.
\(^{14}\) See http://www.itsmarc.com/crs/Auth0679.htm.
\(^{15}\) Available at http://dtd.nlm.nih.gov/publishing/.
\(^{16}\) Also known as the Eprints Application Profile: (http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile).
\(^{17}\) Available at http://www.icacds.org.uk/eng/ISAAR(CPF)2ed.pdf.
\(^{18}\) For details see http://www.eprints.org/software/v3/. 
about their institutional affiliations (and how they have changed over time) and the variant forms of their names. These may be harder to determine without the involvement of the authors themselves. If the Names service is developed beyond a prototype it will be important to be able to accept information from authors so that the accuracy and quality of the information within the service can be improved and maintained.

Authors will also be able to assist with links across systems by recording the identifiers assigned to them in other systems. Multiple registration requirements across a number of name authority systems may create an element of ‘ID fatigue’ for authors. If connections can be made across different services, some of this may be avoided. In the Netherlands, a centralised network of author identities has been created by linking data from research information centres to that of the name authority files in the library system (van Spanje, 2007). It is possible in the future that similar links could be forged between a UK Names service and the researcher information held by the Joint Electronic Submission System which is maintained by the Research Councils in the UK. This would enable these funding bodies to easily see how their funding decisions have been converted into published articles, datasets and other repository materials by the recipients of their grants.

Conclusion
The rapid development of institutional and subject-based repositories has brought with it problems of author identification that have been well-known in the library world for many years. The Names project is attempting to bring decades of library expertise together with currently available technologies and data in order to create a prototype name authority service that will help the new repository landscape to solve one of its most pressing problems: the unique and unambiguous identification of creators of repository content. In the future, the involvement of the authors themselves will be essential to maintain and improve the service.

References


Classifying by Phenomena, Theories and Methods
Examples with Focused Social Science Theories

Abstract
This paper shows how a variety of theories employed across a range of social sciences could be classified in terms of theory type. In each case, notation within the Integrated Level Classification is provided. The paper thus illustrates how one key element of the León Manifesto — that scholarly documents should be classified in terms of the theory(ies) applied can be achieved in practice.

The León Manifesto (http://www.iskoi.org/ilc/leon.htm) argues that interdisciplinary research would benefit from a universal system of document classification that classed each work in terms of the phenomena studied, the theory(ies) and theory types applied, and the method(s) applied.

Classification by phenomena, instead of disciplines, is being experimented with in the Integrative Level Classification (ILC) project (http://www.iskoi.org/ilc/) (Hong 2005). In ILC, main classes of phenomena are represented by letter codes, such as:

<table>
<thead>
<tr>
<th>Letter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>communities</td>
</tr>
<tr>
<td>sf</td>
<td>families</td>
</tr>
<tr>
<td>sm</td>
<td>tribes</td>
</tr>
<tr>
<td>t</td>
<td>organizations</td>
</tr>
<tr>
<td>tn</td>
<td>states</td>
</tr>
<tr>
<td>tr</td>
<td>federations</td>
</tr>
<tr>
<td>tu</td>
<td>supra-national unions</td>
</tr>
<tr>
<td>u</td>
<td>economies</td>
</tr>
<tr>
<td>u5</td>
<td>capital facet</td>
</tr>
<tr>
<td>u5b</td>
<td>labour</td>
</tr>
<tr>
<td>u5q</td>
<td>equipment</td>
</tr>
<tr>
<td>u5s</td>
<td>structures</td>
</tr>
<tr>
<td>v</td>
<td>cultures, technological systems</td>
</tr>
<tr>
<td>vd</td>
<td>pastoralism</td>
</tr>
<tr>
<td>vm</td>
<td>farming</td>
</tr>
<tr>
<td>vt</td>
<td>industry</td>
</tr>
</tbody>
</table>

Theories and methods can be expressed in the form of facets, introduced by the digit indicators 04 and 03 respectively. Thus, the subject “states studied by theory x and method y” will be notated tn04x03y. The theory facet has been the least developed element in this approach. Only a dozen grand theories have been classified (Szostak 2004). The purpose of this paper is to show that it is possible to classify documents employing narrowly focused theories. It will place theories from a variety of social science fields — some disciplinary and some interdisciplinary — into a typology, and develop ILC notation for classifying works using these theories.

1. Theories and theory types
Theories present a challenge to document classification because theory names are ambiguous: the same theory name can refer to quite different sorts of theoretical argument, while quite similar arguments may go under quite different names (especially in different disciplines). The researcher wondering if a particular theoretical argument has been applied to a particular set of phenomena will receive limited guidance if the literature is classified only with respect to theory names. The recommended solution is that works also be classified in terms of a typology of theory types (Szostak 2007, 2008).

Szostak (2004) developed a simple five-dimensional typology of theory types through recourse to one of the simplest classificatory devices: asking the 5W questions, who, what, where, when, and why. These in the context of theory yield more precise questions
(and in each case a mere handful of possible answers), which can be represented as facets 047, 046, 045, 043, 040:

046 Who is the agent? There are two immediate distinctions here: non-intentional (including volcanoes or institutions versus intentional agency (of beings that can act on purpose), each of which can take the form of individual, group, or relationship agency. This can be represented by specific classes, taking their notation from ILC main classes (*extra-defined foci: see Gnoli 2006*) like 046g matter, 046p persons, 046s groups, 046t institutions, 046w artifacts, etc.

043 What does the agent do? There are three broad answers: 043e passive reaction, 043i active action, 043n changes in attitude.

045 Why does the agent do this? With non-intentional agents, action can only be understood in terms of 045b their inherent nature. With intentional agents, scholars can explore the five distinct types of decision-making: 045y rational, 045i intuitive, 045p process (virtue) oriented, 045v rule-based, and 045w tradition-based. For groups and relationships, scholars can also ask how individual preferences are aggregated.

040 Where does the causal process occur? How generalizable is the theory?: there is a continuum from 040o nomothetic (highly generalizable), through 040m half generalizable, to 040k idiographic (situation- or causal-link-specific) theory.

047 When does the causal process occur? Though inspired by the temporal question ‘when?’, the possibilities refer ontologically to directions of change. There are five broad time-paths that a causal process might follow: 047b return to the original equilibrium, 047c cyclical oscillation, 047e movement to a new equilibrium, 047p change in a particular direction, or 047s stochastic/uncertain.

As the discussion above suggests, theories may occupy multiple cells in the typology. This may occur because a theory has changed through time, or because theorists have not provided clear answers to one or more of these questions.

Well-known and consolidated theories, like Classical political economy or Marxism, can be assigned a specific theory notation: sf)04um will mean “families studied by Marxism”, while tn)04um “states studied by Marxism”. Their characterization in terms of theory types can be recorded as a relation of dependence (Gnoli et al. 2007) on the appropriate types 047e 046s 043i etc. On the other hand, works applying innovative, not well-agreed or more narrowly focused theories can be classified only in terms of theory types: u047e046v “economies studied by theories postulating movement to a new equilibrium and institutional agents”. If the search interface is programmed appropriately (Gnoli & Hong 2006), users will be able to search for any work postulating movement towards some new equilibrium (047e): the system will retrieve both documents applying consolidated theories like Marxism (thanks to the relationship recorded in the database between 04um and 047e), and documents applying some other, less consolidated theory implying movement to a new equilibrium.

2. Methodology of this paper
A random selection of works housed in the library of European University Institute in Florence, Italy was used for this paper. A subject search in the online catalogue

---

1. Bracketed notation means a common facet (Marxism as a general theory, by which any phenomenon can be studied) rather than a facet typical of the present class of phenomena (such as a specific theory of the state).
under ‘social science theories’ produced thousands of hits; the first 100 books (temporally most recent) were consulted and those that surveyed several theories in a particular field were selected. Two books were added to the sample because they were located near books being investigated and fit the search criterion. This methodology yielded five books from five distinct human science fields and with a mix of single authorship, co-authorship, and edited volumes.

The strategy for theory notation in ILC is to give each widely-used theory a unique theory code, but also classify each theory in terms of the most common theory type. Works using a less common version of the theory would receive a notation that indicated the deviation from the most common theory type. In some cases, theories do not specify answers along each of the five dimensions; applications of that theory which address that dimension would then receive additional notation for that dimension. The unique notation for each theory is provided below. The discussion highlights the most common theory type, and most typical deviations from this.

3. **Theory of Development** (Larrain 1998)²

**Classical Political Economy (Smith, Ricardo, Malthus)** (or more precisely “economic development according to classical political economy”, and similarly with the following theories): These theories, which formed the basis from which modern economics developed, would be expected to be similar in type to the characteristics identified for rational choice theory in Szostak (2004): individuals, action, rationality, equilibrium, and highly generalizable: u3d047e046p045y043i040u. Certainly, both Smith and Ricardo theorize an eventual equilibrium where all possibilities for growth are exhausted, while Malthus theorized that humanity could never raise average incomes for the masses above subsistence.³ In invoking trade and division of labor as major sources of economic growth, Smith especially sometimes invokes relationship agency, but this would be the major deviation.

**Classical Marxism**: Marx in some passages emphasized technological determinism 046w, and in others stressed the importance of class struggle. The first is a form of non-intentional individual causation, while the latter reflects intentional group agency 046s. The first invokes passive reaction 043e to technological innovations. The second argument also at times assumed inevitability but at other times Marx (and especially later Marxists) stressed active action.

Marx spoke at times of class consciousness, and thus did not neglect attitudes — though note that if the process of class conflict is assumed to be inevitable then these attitudes play no distinct causal role. When Marx assumed the inevitability of historical processes, his theory predicted the move to a new (socialist) equilibrium 047e. When he (or his followers) allowed for historical contingency, then his theory became stochastic 047s. At such times, his theory thus also allowed for intentional individual or relationship agency. While Marx is generally perceived as a generalist 040o, he did at one point caution that the historical process he outlined applied only 040m to Europe (Larrain 35).

---

² While Larrain takes a Marxian perspective, and stresses the role of class conflict in different theories, his discussion notably still provides all of the information needed for classification.

³ Of course, they also speculated on the nature of the dynamic process that would take them to this new equilibrium. Ricardo stressed both forces leading to increased incomes (innovation, trade) and to decreases in income (diminishing returns as more people worked with a fixed supply of land). He might be thought of as describing dynamic processes in particular directions.
While Marx differs along many dimensions from classical political economy, Larrain at least would stress his emphasis on class. Szostak (2004) noted that with respect to group and relationship agency one would have to classify works in terms of how individual preferences were aggregated: democratically, through negotiation, by the exercise of authority. Marx of course stresses power relationships. But the exercise of power is stressed between classes, and thus it is not always clear how decisions are made within classes. It may nevertheless be a useful addition to the typology developed in Szostak (2004) to distinguish theories of power inequalities between groups (or individuals or within relationships) from theories that assume mutually advantageous interactions.

**Classical Theories of Imperialism:** As with the classical Marxism from which most of these theories derived, most writers saw imperialism as a prelude to the inevitable collapse of capitalism, and can thus be characterized as describing the movement to a new socialist equilibrium. But some authors (especially later) stressed that socialist revolution was not inevitable and thus described a stochastic process.

Theories of imperialism stressed the role of large corporations. These supported imperialist policies in pursuit of raw materials and/or markets (different authors stressed different motives. These theories thus stressed individual intentional agency more than did classical Marxism (though Marx also talked about the relentless pursuit of profit by individual capitalists). Note that the corporation is legally an individual, though one might still wonder how exactly corporations reach decisions: writers in the Marxian tradition would stress the exercise of power within the corporation.

With respect to generalizability, early authors tended to argue that the colonized would develop in similar directions as the colonizer, but later authors doubted this.4

**Modernization Theory:** Larrain speaks of three types: sociological, psychological, and economic. All three stress the inevitable movement of poor countries toward the economic, political, and cultural attributes of rich countries. They are all thus theories of change in a particular direction (though some eventual equilibrium might be imagined). Modernization theories stressed the direction of change: sociologists emphasized the breakdown of tradition, increased division of labor, increased importance of merit; psychologists emphasized achievement motivation; and economists stressed the putting in place of institutions and infrastructure that would allow economic growth. All were vague in describing the agents, actions, and decision-making processes driving modernization (though the psychological version arguably usually stressed individuals, attitudes, and perhaps rationality). All assumed generalizability.

**Theories of Worsening Terms of Trade (Prebisch):** Faced with sluggish economic growth especially in Latin America, these theories hypothesized that poor countries would not benefit from international trade as much as rich countries (due either to the fact that technological innovation would reduce the relative value of raw materials, or because rich countries protected local sources of raw materials), and urged states in poor countries to restrict trade in order to encourage domestic industrialization. The technological determinism can be treated as above as non-intentional agency. The emphasis on

---

4. If a document emphasizes real phenomena such as ‘corporations’ this emphasis will be captured within the phenomena facet. If, however, a document hypothesizes theory-specific elements such as ‘colonized mentality’ these must be captured within the theory facet.
the state is a particular sort of group agency; these theories tended to assume that states would act in the interest of their population.

Dependency Theory: These theories built upon but also critiqued the previous theories of worsening terms of trade. Palma had identified three types of these (in Larrain 112): those that assumed underdevelopment in poor countries as inevitable, those that identified problems but eschewed inevitability, and those that stressed how internal class conflicts mediated the effects of international economic relationships. Larrain (114) instead distinguishes highly generalizable theories of inevitable underdevelopment from historical analyses which tend to stress class conflict and see development as a possible outcome. Larrain’s first type is equilibrium and generalizable and tends to stress non-intentional agency. The second type is non-generalizable, stochastic, and tends to stress group agency (and power relations between groups). But the first type also has important elements of group agency: it is often argued that elites in poor countries serve the interests of rich country capitalists rather than of poor country workers.

Both types stress the ‘exploitative’ relationships between rich countries and poor countries. Larrain is critical of the first type for often leaving the mechanisms by which this occurs vague, and compares them to modernization theory in this respect (130). The second type is diverse though it tends to emphasize the actions of classes and states.

4. Theories of International Relations (Burchill and Linklater, 2005)

Most but not all of the theories discussed in this book are general political theories applied to the area of international relations. Some are economic theories likewise applied, hence they take bracketed notation. The focus on international relations would be captured in the classification with respect to phenomena. This field, like many, has intense debates regarding epistemological issues: to what extent, for example do theories ‘create’ reality rather than merely describe this? In the former case, subject components should be represented as phenomena, while in the latter, as theories and theory types. The editors do a good job of distinguishing these types of arguments from the ‘scientific’ questions of how international relations works: indeed they use quite similar terminology to Szostak (2004) in distinguishing philosophical from scientific theory. Nevertheless it is clear that theoretical discussions in the field often conflate the two, and thus that works may often need to be classified in terms of each. Theories in international relations are most clearly distinguishable in terms of the results posited, and thus often provide only limited guidance for classifying with respect to questions other than ‘where.’ The editors stress the lack of homogeneity within the theories that they discuss.

Liberalism: Liberal theory stresses how free trade supports both economic growth and world peace. Recently, the advantages of universal human rights and other ‘liberal’ institutions have been stressed as well. These theories can generally be characterized as outlining change in a [desirable] direction, though some works describe a future equilibrium. Less clearly, these theories generally emphasize how institutions constrain actors. The actors of chief concern are governments themselves, or sometimes individual government leaders.

5. Szostak (2004) discusses how philosophical theories might be classified. Discussions of critical theory and postmodernism are not treated here because they are primarily philosophical in orientation.
**Realism:** Realism can be distinguished from liberalism primarily in being much less optimistic about the possibility of moving away from the present situation of an anarchic contest for power between states. Realist theory of international relations is thus an equilibrium theory. Like many sorts of systems theory, it is vague regarding key causal forces (though individual works may be more precise), though there is a common interest in the limited constraints on state power in the international arena.

**The English School:** This approach seeks a middle ground between the above two.

**Marxian Approaches:** In this context Marxian analysis tends to stress the nature of ‘capitalist forms of production.’ Technological determinism is key, though class relations still important.

**Constructivist Theory:** This stresses how ‘anarchy is what states make of it.’ That is, attitude formation is given a central role. Unlike other theories, the time path is considered less predictable (stochastic), for states have the ability to fashion different futures. Elite groups are often emphasized, though other sorts of agency are possible.

**Feminist Theory:** This sees merit, but also difficulties, in all of the above approaches. Individual works may thus pursue a diverse range of types of theory.

**Green Theory:** Once again the emphasis is on results. Green theorists hope to identify institutions or sometimes attitudes that would encourage better treatment of the environment. It thus seeks change in a particular direction.

5. **Theories of the State (Pressman, 2006)**
6. **Theories of Industrial Relations (Muller-Jentsch, 2004)**
7. **Contemporary Social Psychological Theories Burke (2006)**

8. **Discussion**

This paper has surveyed a handful of research areas in social science, and shown that the theories encountered in each can be reliably classified in terms of theory types and provided with appropriate ILC notation. Such a classification would aid researchers in identifying works that apply a particular theory type to a particular phenomenon.

In some cases, classification is rendered difficult by the vagueness with which certain theories are expressed, or by ambiguity in how a particular theory is interpreted by different theorists. Theory classification should thus act as a tool, and at the same time as a stimulus, for a more clear description and definition of theories. Authors could be encouraged to classify their own theories in terms of theory types and as compared with other existing theories. Author classification could then be overseen by classificationists to ensure that works are placed in suitable places in classification schemes.

The approach outlined in this paper, of classifying works in terms of phenomena studied and theories and theory types (and methods) applied is part of a larger model, which is described programmatically in the León manifesto (cited above). It may be that, at least for some uses, works should also be classified along other dimensions such as local viewpoint (Beghtol 1988), epoch of knowledge (Tennis 2002), application to human activity (Vickery 2008), and disciplines; ILC notation is available in each case. First experiences

---

6. Some twenty different theories discussed in these three books were also classified and provided with ILC notation. The discussions are omitted here due to space limitations. Importantly, it was similarly straightforward to classify theories in these three domains as in the two domains addressed above. These discussions can be found at www.iskoi.org/ilo/montreal.htm.
with indexing bibliographies by ILC suggest that dimensions occur with different frequencies in different domains: in a natural science domain like bioacoustics, method is expressed far more frequently than theory (Gnoli et al. 2008), while the opposite seems to happen in human sciences.

More experimentation is obviously needed to assess the effectiveness and the details of this approach, as well as the expression of theories and methods in domains other than social sciences.

References


Knowledge Organization Pro and Retrospective

Abstract
This paper discusses former and expected general developments in the area of knowledge organization. A distinction should be made between long-term and mid-term developments. In addition to these results, significant past developments and their aspects were discussed at a Vienna discussion panel 2006 and at Bangalore and Juelich panels in 2007. These yield a list of expected or suggested future developments.

1. What is Knowledge Organization
Ingetraut Dahlberg writes: Knowledge Organization is 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 (Dahlberg 2006).

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.

At another point Dahlberg stresses (Wiss-Org 2006): “Henry Evelyn Bliss used the composed term, ‘Organization of Knowledge’ in his two books published 1929 and 1933, respectively, i.e. “The Organization of Knowledge and the System of the Sciences” and “The Organization of Knowledge in Libraries.” However, we also considered the term ‘knowledge order,’ a designation which we had equated with ‘classification’ when founding the German “Gesellschaft für Klassifikation” in 1977. … However, after some discussion we favoured the term ‘Wissensorganisation — Knowledge Organization’ as it allowed a direct translation into English, whereas the term ‘order’ in combination with knowledge may be misleading, because of the verb ‘to order’ (e.g. a service, a product).”

“The concept of ‘organization’ however, as it is accepted in German has a wider range than just ‘order.’ namely ‘planned construction,’ ‘structure,’ ‘forming’ (Wahrig 1975), although this does not apply to some other languages where ‘organization’ is only used for collectivities like associations or unions, so that in such cases, ‘organization’ can only be related to people, not to objects.”

Birger Hjørland remarks: “You make a difference between the phrases ‘organization of knowledge,’ and ‘knowledge organization’ … I have so far not searched systematically for the history of the term knowledge organization, but the term appears in the following work from 1910: Principles of the Science of Organisation as applied by the Knowledge Organisation Bureau, Limited, in its Bureau Encyclopedias”

To reiterate; there is more to be understood by knowledge organization than just organizing; the processes of saving, finding and communicating thoughts can also belong to the practical processes.

2. Knowledge Organization as a Counterpart of Society

Knowledge organization reflects the efforts of man to bring sense to his experience of the world with which he is confronted on a daily basis. He makes the knowledge used in his work, which serves more or less to ensure survival, conscious and accessible. The organized knowledge becomes an extended genetic code which can be called up as needed. Over centuries and millennia, it can be presumed that urgent progress was needed, such as demonstrated in library science, information science and information technology leading to easy, confident usage of knowledge.

However such developments are neither inevitable nor always the same. There are differing needs in knowledge and knowledge technology according to work environment, conclusions drawn from extant knowledge and derived goals. An agrarian society has different needs than a society affected by the new economy and production; trade or fine art have different requirements regarding type of knowledge and how it is utilized. Knowledge and knowledge technology have, thereby, grown historically alongside. Everything we do or think is built on the shoulders of giants, as Robert K. Merton characterizes the quantitative and qualitative development of the scientific knowledge. Phylogenetically between generations, as well as ontogenetically per person, we make experiences through

---

3. Contribution in June 2006 to Discussion-List Wiss-Org. This thread is currently not available.
4. An example for this is the view that humans, environment and cosmos are equally structured, as early schematic anatomical and architectural analogies suggest it.
5. E.G. the syllogism of Aristotle or Ars Magna of Ramon Llull.
6. As for instance with the linguistic relativity thesis by Saphir-Whorf has been worked out.
7. Whereby also different conversions of ideas are justified, like those of ‘Autopoiesis’ or ‘Post-Modernity’.
the perspective of our predecessors, repeat understandings and change them. Knowledge is thereby in constant flux and the respective current chapter conforms to the current requirements. This applies also to the conscious or tacit competence to deal with knowledge. Today can be said that knowledge organization is more broadly demanded for than ever, what in addition, means that it must be able to be mediated and if necessary to be carried actively to specialized technical and software engineers.

3. Knowledge Organization in the Long Run
A compressed look at the historical progress of knowledge organization points to the following noteworthy occurrences. It began with the coding and availability of knowledge. This became manifest with wood and bone cutting and particularly with Stone. Then these were eventually replaced by transportable pottery tablets and representations in bronze. Libraries created ways of searching for this growing inventory. The printing press served to multiply the knowledge carrier. Knowledge was organized according to logical rules, being string systems or hierarchies. Mass storage allowed for broadening of capacity and circulation. Statistical analysis made more organization and selection possible regardless of the large amounts of knowledge carriers. Knowledge was organized according to logical rules, being string systems or hierarchies. Mass storage allowed for broadening of capacity and circulation. Statistical analysis made more organization and selection possible regardless of the large amounts of knowledge carriers. Flexible availability was increased via communication technology. Graphics served for holistic comprehension of content. Then finally we return to the starting point whereby cave painting already used pictures to codify, even if a bit less strictly in relation to what was being said.

The following aspects played a role in these developmental steps that would trigger the development of knowledge organization in the course of history:

**Content**: Beginning with humanities (religion, philosophy), on through science and technology, product catalogues and collectibles, processes (software, computer models) were also worth collecting.

**Sustainability and the public availability**: Initially only a close circle was responsible for access to knowledge. Then came the class of learned and nobility who saw knowledge as their privilege. Then finally knowledge became part of the cultural assets belonging to the people. Today it is in principle everywhere and available to all.

**Persistence**: From stone and burned clay (resistant and permanent) to correctable wax, and then to quickly coded, lightweight paper. Later, punch cards could be copied and electronic form enabled transformation.

---

8. Just like family sozialisation, where values of parents are taken over, there is also the lifelong sozialisation in the working sphere and in science.
9. Thus Stonehenge is by now regarded as an observatory for typical sky constellations to predict seasons.
10. As found in Mesopotamian palace archives.
11. The library of Alexandria seems to have listed the place of origin, the owners and the editors.
12. Of Schedel’s world history already approx. 1,400 copies in Latin and 700 copies in German were printed. Pirated editions followed.
13. E.G. circle representations (7 free arts), trees of the knowledge, decimal classifications.
14. This concerns analogue memory, like microfiche, as well as digital memory, like magnetic tapes or optical memory.
15. E.G. so called multivariate reduction methods, like cluster analysis.
16. E.G. selforganizing maps or cognitive maps.
Coding: It began with pictures, moved on to symbols and eventually led to language and meta-languages. Indexing languages began with ordering, then subdividing (decimal), to pre- and post-coordination. Facet classification, links and roles enriched the power of coding. Networks and clusters of codes introduced a quality of context and transition.

Processing: It began with filing, later giving way to subsumption (ordering, classification, organization). More sophisticated were polyhierarchies and keywords in context. The distinction between object, concept and notation gave more expressiveness. Then language-processing and logic reasoning was introduced. Now statistical calculations help in reduction and patterns.

Organizing: Rules for title and alphabetic cataloguing were followed by generality of universal classification systems. Rules for controlled terms were followed by metadata recommendations and ontologies for documented types. Gateways and interoperability are examples of recent efforts.

4. Future Mid-Term Expectations from Panel Discussions
At the German ISKO Conference 2006 (International Society for Knowledge Organization) in Vienna, a panel on the future of knowledge organization was convened (Ohly 2008).\textsuperscript{17}

Gerhard Rahmstorf indicated that knowledge, among other things, would grow through research activities, that it is complex and that it must be computer processable. He sees recourse to the language as still necessary. Word meanings and sentences must be clarified and representable by computer technology. The composition of a knowledge base then encompasses the analysis of sentence forms and storage of sentences. Appropriate representation can be presented in tables. Questions can thereby be answered.

Winfried Schmitz-Esser realized that knowledge as known until now in the sense of ‘shelving order’ is no longer sufficient. Digitalized sources in large qualities, with heterogeneous text and in various languages need more organizational means than was previously necessary. A world model is needed that integrates means of organization and that can also detect and process knowledge in texts. Non-language materials must also be integrated and the results presented for the user as comprehensible. In the future it should be possible to track down and evaluate without the help of librarians.

Gerhard Budin stated that Knowledge Organization nowadays is an interdisciplinary effort to engineer shared knowledge. This comprises cognitive, epistemic, communicative, and automatic knowledge representation, creation and processing. One predominant requirement is concept and terminology work.

The view of Michael Nentwich is that information seekers — as well as scientists — are too modest in their searching demands and that subject-specific portals offer worse quality than the existing specialized databases. This includes community driven blogs, social bookmarking, etc. as well. Such half-solutions should at least work against special solutions. Knowledge organization specialists create something like a Google 2.0 when semantic intelligent solutions are sought.

The ongoing discussion stressed — more than ever — the broad demand in knowledge organizations, meaning that it must be also communicated to experts and software engineers.

\textsuperscript{17} See Chap. 6 In: Ohly et al. 2008.
The IKONE conference (International Conference on Future of Knowledge Organization in the Networked Environment) 2007 in Bangalore can generally be summarized as: Knowledge organization is needed worldwide especially in connection with technological enhancing of storage and distribution means. Knowledge Organization is no longer a domain of librarians but also of practitioners, knowledge engineers, and knowledge managers. In connection with the Internet and full text the field of knowledge organization must be seen much more broad and flexible than classification or thesauri systems, such as ontologies. I will enlighten this by stressing out the contents of some selected talks.

In the inaugural session Prasad Bhaarat Ram restated information supply as “Give me what I want, not what I ask for”. Means for that might be procedures that deal with misspellings, spam ranking, user models. Maxmilian Stempfhuber stressed out that aggregation of data resources in portals requires special treating of heterogeneity with respect to user demands.

Kavi Mahesh introduced into the creation of a ‘situated’ knowledge organization scheme for various demands for knowledge management in enterprises. To achieve balance between compositional semantics and nominals certain proliferation principles must be applied: similarity, specificity, opposition, and unique axis. According to Madan Mohan Rao knowledge management should not only focus upon IT. He stated that successful knowledge management practices can be facilitated by adequate access to knowledge management tools, user-friendly work-oriented content, communities of practice, a culture of knowledge, learning capacity, a spirit of cooperation, commercial and other incentives, and carefully measured capital investments and returns.

A concluding panel moderated by Peter Ohly named as biggest challenges in the next five years: Construction of virtual Knowledge Organization schemes (Mahesh), enhancements in education in knowledge organization (Khoo), understanding “human needs” and open modeling that incorporates even non-standardized material (Malone), building methodologies for knowledge organization (Stempfhuber).

A conference on Scientific Communication of the Future in Jülich 2007 centered more or less around similar topics\(^\text{18}\). Emphasis was given on eScience-Tools, collaborative techniques, information networking and open access. The outcome yielded in diversification of knowledge and types of knowledge communication distinguishing between mainstream knowledge and ingenious knowledge creation.

A special discussion panel focused on quantitative aspects of science output, general frameworks for science and research and science as mass phenomena. Individual statements were:

We have too many data — primary and even secondary data. Thus we should not collect too many but carefully select them. The investigation into information acquisition in scientific work is important but too less reimbursed by the funding agencies. Information questions from Information Science migrate to the specialized sciences, but document management rests as a primary function for librarians — but on a high know-how standard. The science of the future requires collaborative communication techniques, but leading research will still require individual efforts. Mass indicators for science efforts must be changed to qualitative examinations where trust plays an important role.

To give a general summary, knowledge organization must realize the increasing diversity of knowledge and communication styles together with public training in organization tools and communication techniques.

**Conclusion**

If lessons are to be learned from the past, then definitely the following. Whatever was new was almost always supplementary, not substitution. Sometimes detailed, and sometimes general information was in demand. Recall (extension) and precision (intension) alternated. Input (documentation) and output (information) mutually determine one another. Durability and flexibility are constant opponents. Self-serving interests in information conflicts with sustainable information availability.

Nowadays knowledge must be compiled and evaluated by many persons, and that means more, faster, and more flexible knowledge processing. Compatibility is not sufficient, instead heterogeneous information needs to be transferable. New object forms — in addition to texts — need be considered. Information must be offered as multi-disciplinary and user-oriented. In general a knowledge organization literacy has to be created since new medium technologies require also new handling skills.

**References**


Wiss-Org 2006. “What is Knowledge Organization”. Contributions to the discussion lists wiss org@bonn.iz-soz.de and isko-l@lists.gseis.ucla.edu in Aug. 2006.
Knowledge and Trust in Epistemology and Social Software/Knowledge Technologies

Abstract
Recent developments in so-called knowledge technologies, social software, web2.0 or even web3.0 applications as new classes of information and communication technologies (ICT) have raised not only societal, political or economic hopes and expectations, but also epistemic ones. As in the beginnings of the internet, utopias are developed about a collective space of knowledge and a networked, open or even democratic organization of knowledge. The key features of these technologies are social networks or trust relationships between users and their exploitation for knowledge assessment. Instead of reinventing the wheel, software developers might benefit from taking a look at considerations on knowledge, sociality and trust in epistemology and related disciplines. In return, theoreticians of knowledge will surely benefit from paying closer attention to new technological developments. Accordingly, the aim of this paper is twofold: The primary goal is to identify connections between notions of trust and knowledge inherent in these social software/knowledge technologies on the one hand and theoretical conceptions of knowledge and trust on the other. Moreover, having identified various starting points, I argue for an intensified intellectual exchange between different theoretical approaches to knowledge as well as between those theoreticians and ICT developers.

Introduction
Recent developments in so-called knowledge technologies, social software, web2.0 or even web3.0 applications as new classes of information and communication technologies have raised not only societal, political or economic hopes and expectations, but also epistemic ones. As in the beginnings of the internet, utopias are developed about a collective space of knowledge and a networked, open or even democratic organization of knowledge. An analysis of the glut of these new social software/knowledge technologies launched every week reveals that the key features of these technologies are linked to notions of, respectively, implementations of social networks or trust relationships. For instance, Twine\(^1\), a new semantic web application launched recently at the Web2.0 Summit in San Francisco promises its users “a smarter way to share, organize, and find information with people they trust.” Moreover, “Twine helps you better leverage and contribute to the collective intelligence of your network”. In these applications, the notions of social networks or trust relationships between users of a system are associated with the evaluation of the quality of knowledge organization, sharing and retrieval. The efficiency of knowledge processing is supposed to be increased by exploiting the collective intelligence of social networks bound together by statements of trust. Based on such observations, I argue that the concept of knowledge employed in these social software/knowledge technologies fundamentally depends on the concept of trust. Yet, notions of trust are not only rendered important in social software/knowledge technologies, but they have also received growing attention in epistemology, i.e. the philosophical discipline concerned with the process of knowing and criteria for knowledge. An epistemological account of trust would have to analyze the function of trust for knowledge and vice versa. Within epistemology, the field of social epistemology is concerned with the social aspects of knowledge (e.g. Schmitt 1994, Goldman 2003) or even the thoroughly social nature of knowledge. 

\(^1\) Please consult http://www.twine.com/about for further information and also for the quotes in this paragraph. [Last visit: 27.02.2008]
knowledge (e.g. Kusch 2002) as thus seems to be most promising philosophical framework for analyzing interrelations between trust and knowledge.

**Trust in Epistemology**

In her article on trust in the *Stanford Encyclopedia of Philosophy*, Carolyn McLeod (2006) states that “trust is an attitude that we have towards people we hope will be trustworthy”, however, as the word *hope* indicates we cannot be sure that the one we trust really will prove to be trustworthy in the end. Thus, trust necessarily involves the risk of being let down and the main philosophical question about trust is, under what conditions trust is warranted. Even though reflections on trust and related topics such as (cognitive) authority and testimony have a long tradition in philosophy, the role of trust for knowledge as well as the social aspects of knowledge have long been neglected. It was not until quite recently that social epistemology has been established as a philosophical field\(^ \text{2} \), but also that trust as an epistemological notion has gained importance.

The notion of trust attracted notice of several philosophers in the beginning of the nineteen-nineties and many of the most relevant publications date back to these days. For instance, Annette Baier, a feminist philosopher, has worked extensively on the role of trust especially in moral philosophy and ethics. She focuses on the role of trust for decision-making and on the function of membership within moral communities (Baier 1991a, Baier 1991b). However, the notion of trust was not only embraced in ethics and moral philosophy, but it was soon introduced to also account for communicative and interactive processes in science and epistemic communities more generally. Also in 1991, John Hardwig’s published “The role of trust in knowledge” in the Journal of Philosophy and by this introduced the notion of trust into a rather traditional epistemological discourse. As indicated by the title Hardwig tries to assess the function of trust for knowledge creation in science. He argues that taking a closer look at scientific practise, we have to rebut the antithetical conceptualization of trust and knowledge still prevalent in epistemology, which states that “[w]e can not know by trusting in the opinion of others; we may have to trust those opinions when we do not know” (Hardwig 1991, p. 693). He notes that in epistemology as well as in philosophy of science, it has been almost in unison assumed that knowledge rests on evidence and not on trust. If knowledge were to rest on trust, it would have to be partly blind, which seemed to be an unacceptable premise.

However, Hardwig argues that analysing current scientific practise even in the most renowned disciplines, casts a new light on the relationship between knowledge and trust. “Modern knowers”, he states, “[…] cannot be independent and self-reliant, not even in their own fields of specialization” (Hardwig 1991, p. 693). His analysis departs from the observation that the majority of research is nowadays conducted in teams and he presents two examples of major scientific achievements in physics and mathematics as case studies in support of his claims. Cooperation in science is supposedly needed to overcome time restrictions on the one hand and to handle the rising specialization in science on the other. Due to this high specialization, scientists do not only lack the time to perform every subtask of their research on their own, but mostly they also lack the necessary expertise in the respective area of research. As a consequence, in scientific co-operations scientists

---

\(^{2}\) Interestingly, the term *social epistemology*, does not originate from philosophy, but has been coined by Egan and Shera (1952), who have stressed the importance of epistemology for library and information science.
have to trust the competency and the honesty of their colleagues. Moreover, they have to rely on what Hardwig calls the “adequate epistemic self-assessment” of their peers, i.e. their ability to realistically assess their own competencies and levels of expertise in the areas of concern (Hardwig 1991, p. 700). Thus, in order to successfully operate in science, scientists need to assess their colleagues not only epistemically but also morally. As a consequence, Hardwig emphasizes the necessity to consider and combine epistemology and ethics when reasoning about knowledge.

To put it in a nutshell, Hardwig (1991) argues that trust is epistemologically even more fundamental than evidence such as empirical data or logical argumentation, because one needs to trust these pieces of evidence and their providers to actually use them at all. Thus, the trustworthiness of members of epistemic communities is fundamental to all scientific endeavours and represents the groundwork of (scientific) knowledge creation. Hardwig realizes that these insights might have far-reaching implications for epistemology and philosophy of science and could require fundamental changes in the prevalent models of scientific practise and epistemological concepts of knowledge. However, he explicitly addresses these thoughts to a rather traditional philosophical and epistemological audience and distances himself from more radical approaches, such as the Strong Programme in the sociology of scientific knowledge (SSK) or the works of Lorraine Code (1987) and Michael Welbourne (1986). This reluctance towards these approaches possibly stems from a rash and undue fear of relativism, which might better be abandoned when trying to understand the relationship between knowledge and trust for collective knowledge creation in science and daily life for several reasons. First of all, as will be outlined in the next section, both approaches towards knowledge, one stressing its contextuality, situatedness and plurality of knowledge forms, the other adhering to universal knowledge claims, have already made their way into technological artefacts and find their technological counterparts for instance in local and global trust metrics (Massa & Avesani 2007). Secondly, I argue that different approaches towards understanding knowledge might unearth different important aspects of knowledge and the practices of its creation and that researchers of each type will probably benefit from examining their proclaimed opponents’ findings more thoroughly. This is all the more true since it has been convincingly argued that philosophers of science and sociologists of science share many basic assumptions (Solomon 2001). I will come back to possible connection points between social epistemology on the one hand and sociology of scientific knowledge (SSK) and science and technology studies (STS) especially with respect to analysing ICT and information systems at the end of this paper.

Trust in Social Software/Knowledge Technologies

Now, as noted before, the notion of trust is not only of rising relevance in the field of philosophy, and especially (social) epistemology, but also receives growing attention in recent trends in software development. In the following I will briefly portray some of the works of Paolo Massa and his colleagues, whose research can serve as a prime example for the import of notions of trust of into social software/knowledge technologies. More specifically, they technically implement notions of trust into so-called recommender systems (RS). RSs suggest new items to users, which he or she might like and are often used.
in commercial websites such as amazon.com. In an experimental analysis of data obtained from an internet website deploying such RSs, Paolo Massa and Bobby Bhattachasjee demonstrated that classical RS techniques have several shortcomings, such as sparseness (sparsity of useful information for existing users), cold start (difficulty to generate recommendations for users' who have just registered) and vulnerability of system correctness to attacks (Massa & Bhattachasjee 2004). In the following I will concentrate on the so-called cold start problem.

“Bootstrapping” in RSs is a term for procedures to meet the cold start problem. When a new user enters a system the system does not “know” anything about this new user and this ignorance makes it difficult to generate appropriate recommendations for her. To counteract this problem, traditionally new users were asked to rate a few items so that the system can “learn” something about the user in order to provide personalized information on interesting items for her. However, especially in large databases necessary correlations are scarce and thus, this procedure often turns out to be quite ineffective. In consequence, Massa & Bhattachasjee (2004) develop an algorithm for “Trust-aware Recommender Systems” and argue that the before mentioned problems “can effectively be solved by incorporating a notion of trust between users into the base CF system” (Massa & Bhattachasjee 2004). The difference between traditional RSs and trust-aware RSs is quite simple: “[w]hile traditional RSs exploit only ratings provided by users about items, Trust-aware Recommender Systems let the user express also trust statements, i.e. their subjective opinions about the usefulness of other users” (Massa & Avesani 2006). This seemingly minor change proves to be highly effective to remedy the shortcomings of traditional RSs especially with respect to the cold start problem because “it is able to exploit trust propagation over the trust network by means of a trust metric” (Massa & Avesani 2006).

The work of Massa and his colleagues is not only interesting for the development of trust metrics and their empirical testing, but also for its reflection of underlying premises as well as possible societal and cultural consequences of different technological solutions. In an empirical study with data from the Epinions.com community, they do not only compare different trust metrics in their handling of controversial users, but also reflect on the underlying societal views and possible consequences of these different metrics (Massa & Avesani 2007).

Epinions.com is a web site, where people can publish reviews they have written about a variety of products and these reviews can be rated. The goal of Epinions.com according to their self-description is to help “[…] people make informed buying decisions. It is a premier consumer reviews platform on the Web and a reliable source for valuable consumer insight, unbiased advice, in-depth product evaluations and personalized recommendations.” Moreover, users can assign binary trust statement to other users, indicating whether they trust or distrust them, which leads to webs of trust. Controversial users are users that receive diverging trust statements from the other users of the community, i.e. some of the users trust them while others express their distrust in them. Trust metrics are techniques for answering questions such as “Should I trust this person?” in

---
4. CF systems are collaborative filtering systems, which automatically generate recommendations based on user opinions only. This technology is most frequently employed in recommender systems (Massa & Avesani 2006).
5. For further information and this quote please consult http://www99.epinions.com/about/. [Last visit: 27.02.2008].
virtual communities and in this inquiry they tackle the philosophical question of whether trust is warranted in a certain situation. Massa & Avesani distinguish between global and local trust metrics. While “[g]lobal trust metrics assign to a given user a unique trust score, the same independently of the user that is evaluating the other user’s trustworthiness […], a local trust metric provides a personalized trust score that depends on the point of view of the evaluating user.” (Massa & Avesani 2007, p. 2).

What is important is that controversial users are valued very differently in the two different metrics. Local trust metrics explicitly stress and appreciate the individuality and situatedness of every trust statements and state that controversial users by definition do not have a global trust value for the whole community. By contrast, global trust metrics suggest a fictitious consensus between users by calculating an averaged trust value for each user. Through this process, the controversial user, who is considered to be malicious anyway, is statistically eliminated. Moreover, Massa & Avesani (2007) argue that these different types of trust metrics do not only have underlying assumptions about the value of those users and about deviation from the mean — or norm — more generally. Rather, they also retroact on cultural and societal values, i.e. about how to deal with minority views and thus can and should be linked to critical approaches in epistemology, especially feminist epistemology. It should be noted that this acknowledgement of mutual interactions between technology on the one hand and societal and cultural values on the other also puts their work into the wider context of science and technology studies (cf. van House 2003) and certain critical approaches within technology assessment (e.g. Schot & Rip 1997). To be sure, Massa and Avenani (2007) conclude that both types of metrics do have their vantages and their respectively preferred areas of application, however, from an epistemological point of view, they can also be related to theoretical differences between notions of knowledge that stress the situatedness and contingency of all knowledge claims and advocate pluralism of knowledge and those that rather adhere to universal knowledge claims. A closer analysis of these underlying assumptions and their relation to epistemological notions of knowledge, sociality and trust thus seems to be rewarding.

**Conclusion and relation to further research**

Obviously the concept of trust is not only needed in epistemology to account for and describe the relationships between scientists and their relevance for the creation of scientific knowledge, but also for the creation of knowledge on a daily basis. Trust, it seems, is indispensable for knowledge creation and its implementation in recommender systems to improve the quality of information supply will even gain importance in the next years as the use of social software/knowledge technologies will supposedly rather increase than decrease. Assessing the quality of information, deciding whom to trust and whom to distrust is not limited to information obtained from the internet. However, it becomes all the more obvious in an environment in which information can be exchanged with high speed over long distances, enormously increasing the amount of interaction with people we do not know personally but whom we have to trust — or decide to distrust — nonetheless. Taking these developments into account, it becomes all the more obvious why the notion of trust is becoming more and more important in different fields.

I argue that social epistemology as well as science and technology studies (STS) provide a rich framework for analyzing interrelations between conceptions of knowledge, sociality and trust and technological developments. Within these two domains informa-
tion systems and ICT have been under investigation for quite a while. However, despite some exceptions (e.g. van House 2002), the intellectual exchange between these different theoretical lines of thought has been quite hesitant so far. I therefore want to conclude this paper asking for more intellectual exchange between theoreticians of knowledge and software developers on the one hand. On the other hand, also within theoretical discourses, less reservation and scepticism between different theoretical orientations might lead to cross-fertilization and can only be beneficial for understanding a phenomenon as complex as knowledge.

References

6. For further information, please consult the very elaborate reviews of Don Fallis and Nancy van House’s in Annual Review of Information Science and Technology (ARIST). While Fallis (2006) gives an overview on socio-epistemological analyses of information systems, van House (2003) displays the work on information systems conducted in the field of STS.
D. Grant Campbell — University of Western Ontario

Derrida, Logocentrism, and the Concept of Warrant on the Semantic Web

Abstract
The highly-structured data standards of the Semantic Web contain a promising venue for the migration of library subject access standards onto the World Wide Web. The new functionalities of the Web, however, along with the anticipated capabilities of intelligent Web agents, suggest that information on the Semantic Web will have much more flexibility, diversity and mutability. We need, therefore, a method for recognizing and assessing the principles whereby Semantic Web information can combine together in productive and useful ways. This paper will argue that the concept of warrant in traditional library science, can provide a useful means of translating library knowledge structures into Web-based knowledge structures. Using Derrida’s concept of logocentrism, this paper suggests that what while “warrant” in library science traditionally alludes to the principles by which concepts are admitted into the design of a classification or access system, “warrant” on the Semantic Web alludes to the principles by which Web resources can be admitted into a network of information uses. Furthermore, library information practice suggests a far more complex network of warrant concepts that provide a subtlety and richness to knowledge organization that the Semantic Web has not yet attained.

Introduction
The concept of warrant in traditional library science has been a cornerstone of our understanding of, and our training in, the major tools of subject access, including the Library of Congress Subject Headings, Library of Congress Classification, and the Dewey Decimal Classification. As the articulation of the intellectual basis for the inclusion of entities in a formal ontology, it offers a suggestive means of exploring how LIS traditions can play a role in emerging Web technologies. When we compare the principles of warrant in the two domains, we find that library articulations of warrant principles center on the decisions of a central repository, or of those responsible for large, ubiquitous tools for subject access, tools that give a clear and distinct shape to library collections. On the Semantic Web, questions of warrant surface at the individual level who would set individual preferences: the warrant involved is the decision of which namespaces to admit as trusted and useful domains of information use.

Furthermore, library articulations of warrant are a curious combination of the formal and the informal, the written and the spoken, subjects for ongoing debate, deliberation and slow consensus-building. Semantic Web standardes, on the other hand, take these urgent questions and turn them into formal, machine-readable standards. This paper uses Derrida’s theories of logocentrism to explore these two tensions: between centralized and decentralized questions of warrant, and between warrant as a set of written and oral traditions and warrant as programmable code.

The Concept of Warrant in Library and Information Studies
Svenonius (2000) describes warrant as “the specifying criteria for term selection” in a subject language (135). While this definition is certainly true, Beghtol’s definition (1986) provides a more explicit reference to the status of warrant as a two-stage “authority”: first to justify and subsequently to verify decisions about what classes/concepts to include in the system, in what order classes/concepts should appear in the schedules, what units classes/concepts are divided into, how far subdivision should proceed, how much and where synthesis is available, whether citation orders are static or variable and similar questions. (110)
Warrant, according to Beghtol, is an authority that underlies an entire network of practices and decisions that occur not just during the creation of the subject access scheme, but also during its subsequent use and subsequent revision. Beghtol isolates four distinct concepts of warrant:

1. **Literary** warrant, in which the subject scheme is shaped by a pre-existing literature, either in a specific library or a specific bibliographic domain (112);
2. **Scientific** or philosophical warrant, in which the scheme reflects a prevailing scientific or philosophical consensus about the nature of knowledge in its various fields (115);
3. **Educational** warrant, in which the scheme reflects the practical needs of educational institutions (117);
4. **Cultural** warrant, in which the scheme reflects prevailing cultural values and assumptions (119).

Svenonius, by contrast, isolates three conceptions of warrant: literary, use and structural. Use warrant reflects the perceived needs of information users (135), while structural warrant allows for the inclusion of terms that may not be in common practice, but which provide a means of collocating concepts in a useful way (136).

Despite their different formulations, both Beghtol and Svenonius agree that subject languages rely on the useful exploitation of semantic and syntactic dimensions of subject access, and that warrant provides a means of understanding how the semantic dimension—the selection, arrangement and normalization of terms within a subject domain—manifests itself over time. Warrant can be viewed as a library’s ongoing commitment to analyze, justify and revise the semantic elements of its access systems.

Warrant, then, introduces a new dimension to the relationship between a subject access scheme and the documents that it organizes. Classification schedules typically relate to library collections in a way similar to Ferdinand de Saussure’s distinction between langue and parole: between the sum of all that could potentially be expressed by a language, and the sum total of language “acts” that comprise what is actually expressed. (Saussure 1986, 29) But if we include warrant in this relationship, we realize that warrant criteria monitor and shape the relationship between the system and its documents, between the class numbers we could assign and the class numbers that we actually assign.

But how does warrant actually do this monitoring and shaping? This is a difficult question, largely because warrant appears in many different forms: in the instructions, manuals and formal procedures that accompany our subject access tools, in the formal written policies that underlie the tool’s use in a specific organization, in the inculcation of classification principles in library schools and training sessions, in the deliberations and procedures for revising subject tools, in the various ways in which the tool is discussed, debated, revised, queried, challenged, defended and rejected. But warrant monitors and shapes subject access on at least three distinct axes of implementation:

- the use of the tool to classify existing documents, in accordance with the rules laid out in the accompanying manuals;
- the decision to reclassify documents, in accordance with a change in cultural perspectives or user needs;
- the decision to change the classification structure itself, in accordance with changes in cultural perspectives or user needs.

Along all three axes, libraries and information organizations of all types use principles of warrant to scrutinize both the success with which they have implemented a specific
subject structure, and the degree to which that underlying structure—be it a reflection of 
existing literature, philosophical or educational consensus, cultural values or user pat- 
terns—remains valid.

**Analogies to Warrant on the Semantic Web**
The Semantic Web is a visionary project spearheaded by Tim Berners-Lee and carried out 
largely under the auspices of the World Wide Web Consortium. It rests on a principle of 
machine-understandable data: systems that would enable an intelligent agent to perform 
logical inferences upon data retrieved, to get specific answers to specific questions, to 
mine databases, and to discover relevant resources in diverse knowledge domains. These 
principles manifest themselves through a layered architecture, in which sophisticated 
ontologies such as OWL (Working Ontology Language) rest upon broader stores of meta-
data such as RDF (Resource Description Framework), all of which is encoded in XML.

Like many visions, this one is ambitious; unlike many visions, this one is focused on 
standards:

> Progress is measured on a scale that starts with the conception of an idea, which works through its ac-
> ceptance as a common technology and becomes an interoperable standard, and ends with a global market 
based on it. … [The Semantic Web] absolutely needs the interoperable infrastructure that only global 
standard protocols can provide”. (Berners-Lee 2003, xi)

The W3C promotes these standards, in hopes that their widespread adoption will lead 
to a new transformation in Web culture, a transformation that will broaden the range of 
Web applications available to us, and in so doing dramatically reinvent the relationship 
between humans and computers.

Not everyone believes that the Semantic Web will achieve such widespread adoption: 
some believe, along with Clay Shirky, that the Semantic Web is good for creating syl-
logisms, but not much else (Shirky 2005). But the library community, with its well-de-
veloped access tools, has begun to look to the Semantic Web with considerable interest. 
Greenberg (2007, 203) notes the close analogies between library services and Semantic 
Web activities. Harper & Tillett (2007) suggest that the Semantic Web could serve as a 
means of moving these highly-structured subject access systems into Web environments, 
where they could greatly enhance knowledge accumulation and sophisticated retrieval 
(47), while Campbell & Fast suggest ways in which bibliographic description in academ-
ic libraries could be transformed by closer ties with Semantic Web data (2004).

If libraries do play a part in the Semantic Web, what role, if any, will concepts of war-
tant play? Initial indications suggest that warrant will be of pressing importance. While 
the Resource Description Framework provides a useful and rich means of creating meta-
data about Web resources, some allowance will have to be made for the fact that multiple 
communities will be creating and using this metadata, and this will cause problems when 
search agents attempt to aggregate Web data from multiple sources. Most Semantic Web 
applications will rely heavily on an ontology: “a formal, explicit specification of a shared 
conceptualization” that offers to both humans and web agents “a shared and common 
understanding of some domain that can be communicated among people and application 
systems” (Fensel et al. 2003, 11). Like the tools of libraries, ontologies will determine not 
just the terms to be admitted, but the relationships among them:

> An ontology establishes the things that a system can talk and reason about. This means the vocabulary, 
but … there’s more to it than just a collection of words and names—the terms have logical relationships 
to each other that need to be specified, and this in turn means that any ontology system must adopt some 
variety of logic, either formally or informally. (Passim 2004, 133)
Ontologies, like classification systems, enumerate the entities in a domain, assign them names, and establish relationships among them: all activities which principles of warrant both initially justify and subsequently verify (Beghtol 1986, 110). Ontologies, like classification systems, are presumably products of warrant. Unlike most traditional classification systems, ontologies are designed for machine readability and machine implementation. And unlike most classification systems, ontologies are designed to extract and aggregate data at a very granular level, from multiple sources.

What matters, then, is not the structuring principle of each individual ontology, but the principle by which data from each ontology is combined together. If the closest analogy to a classification system is an ontology, perhaps we can find warrant at work in the mechanism by which an ontology is invoked: the XML namespace. All Semantic Web standards rely on the principle of the XML namespace: the means of avoiding “collisions” between elements with the same name but different meanings by expanding them with a prefix that designates their origin, or “namespace” (W3C 2006).

By limiting and discriminating between different communities’ use of similar element and attribute names, XML namespaces provide a rough, machine-readable analogue to the concepts of warrant used in classification systems. But of course, the concept of warrant mutates considerably in the Semantic Web context. Instead of articulating the intellectual or institutional authority that justifies the inclusion or exclusion of a category in a classification system, the XML namespace provides a machine-readable manifestation of that authority, which is used to include or exclude Web resources in three main spheres of activity:

- Resource aggregation: namespaces, as well as crosswalks that map between similar elements in different namespaces, are used when harvesting resources across different domains, to ensure that only those resources that meet the search criteria are found;
- Data extraction: namespaces provide the means of identifying with accuracy the relevant fields and attributes of databases, enabling agents to extract data from the deep web with precision and selectivity;
- User-Resource interaction; as libraries discovered long ago, effective information use depends on translating the user’s needs into the language of the information system, a process that is frequently iterative and interactive in nature.

XML namespaces, therefore, provide a useful means of connecting library systems of metadata, classification and thesaural terms to a Semantic Web environment; namespaces for the Dublin Core, for subject heading lists, and for classification categories provide the means whereby the rigour of the intellectual structure of a given field of knowledge can be joined with the power of the search agent, allowing only those resources into a search set that satisfy the demands of this structure. While “warrant” in library science traditionally alludes to the principles by which concepts are admitted into the design of a classification or access system, “warrant” on the Semantic Web alludes to the principles by which Web resources can be admitted into a network of information uses.

But what effect is produced by the movement from human understanding to machine understanding, and from the movement from warrant principles that govern a classification scheme to the warrant principles that identify appropriate namespaces? The movement of computer technology into the process of reasoning and inference has raised suggestive issues for knowledge organization, and in particular for the complex judgments that regulate and administer the ongoing adaptations of warrant principles for large-scale
information access tools. To analyze those questions, we need a theory that specifically addresses oppositions between the artificial and the real, between the outside and the inside, and between artifice and spontaneity.

**Derrida and the Logocentric View of Knowledge**
Since its first appearance in 1967, Jacques Derrida’s *Of Grammatology* has raised powerful questions about the ways in which we implicitly rank various forms of discourse, various forms of expression. Derrida coins the terms “logocentrism” and “phonocentrism” to characterize what, for him, is Western culture’s lamentable tendency to devalue writing in favour of speech. Derrida argues that linguistics, as pioneered by Ferdinand de Saussure, proceeded from a phonocentrist bias which led Saussure to treat the spoken word as the primary signifier of the signified, with the written word as the exterior copy: “the voice, producers of the *first symbols*, has a relationship of essential and immediate proximity with the mind. … The written signifier is always technical and representative” (Derrida 1976, 11). For Saussure, Derrida argues, the purity of spoken language is always fighting the corruption and debasement of writing. Derrida maintains that this battle has been waged ever since Plato’s *Phaedrus*, and that the battle was always already lost. The supposed preeminence of speech over writing relies on a series of binary oppositions that were suspect from the beginning: presence and absence, signifier and signified, exterior and interior. As these oppositions collapse under Derrida’s analysis, so too do the logocentric assumptions of language as the articulation of presence.

Derrida’s deconstruction of phonocentrism may seem far removed from concerns of warrant on the Semantic Web. But the Semantic Web, and indeed the Web as a whole, have forced us to reopen many questions in our culture about what constitutes a “real” communication as opposed to an artificial one. Dave Weinberger (2002) challenges many of our assumptions about the Web, especially the assumption that it has no space: the Web, he claims, is an environment of lived space rather than measured space, and one in which human contact is the norm rather than the exception (32). And at first glance, the Semantic Web seems to defy the logocentric bias which, according to Derrida, pervades Western society and intellectual thought.

**Logocentrism and the Semantic Web**
Berners-Lee’s grandiose vision of the Semantic Web rests upon an “interoperable infrastructure” of “global standards,” just as the Web we know today rests upon the widespread adoption of the URL, the Hypertext Transfer Protocol, and HTML. For Berners-Lee, the Semantic Web’s suite of standards, including RDF, RDF-Schema, OWL and XML Namespaces, holds a similar relation to ontologies and Web objects that warrant principles occupy in relation to classification schemes and classified documents. In both cases, the relationship between document and ordering scheme rests upon a generative principle of justification and verification that governs, not just how the ordering scheme is created, but how it is subsequently interpreted and revised. The Semantic Web, however, takes these diffuse principles and articulates them as formal, machine-readable standards. In so doing, the Semantic Web appears to have taken the acts of computer programming, standards creation and machine processing—acts which rarely appear to the popular imagination as primary creative acts—and placed them at the very center of an emergent Web information environment. At the heart of Web interactions lie documents in the form of standards and protocols that govern the behaviour of Web agents. Just as warrant is
the principle that lends coherence to our semantic knowledge structures, Semantic Web protocols are designed to lend coherence to our interactions on the Web: by resolving ambiguities, aggregating facts from different sources, and finding meaning and pattern in data stores too massive for humans to interpret without assistance.

However, before we start hailing the Semantic Web as the world’s first truly post-structuralist information architecture, we should remember that these standards are designed to remain largely out of sight. Indeed, RDF in its XML form is so complicated and difficult to read that its widespread adoption depends largely on the design of applications that can render it all but invisible. Much of the Semantic Web rhetoric suggests that Web agents would adopt a whole new incarnation of “presence”: they would propose solutions to current problems (Berners-Lee & Hendler & Lassila 2001) or prompt users with suggestions (Passim 2004, 8). And because the data structures and underlying ontological assumptions of the various SW standards would remain largely out of sight, they would remain largely outside the user’s awareness.

Conclusion: The Library Community, Warrant, and the Future of the Web

The prospect of using tools such as DDC, LCC, UDC and other schemes and thesauri as a basis for the Semantic Web is certainly intriguing, and certainly deserves further study. But the library community has something to offer besides its classification systems, based as they are upon physical documents in centrally-organized physical repositories. Greenberg suggests that libraries and library systems grew out of a collaborative spirit (2007, 207), and this collaborative spirit comes in particularly handy around issues of warrant. The binary nature of computer data forces automated systems, even those with sophisticated inference capabilities, into a naive position in relation to their retrieval practices and their knowledge structures. Because the namespace principle attempts to resolve ambiguity to heighten the performance of search agents, it forces the agent to take seriously a series of oppositions that librarians have learned, through experience, to treat with some irony: the opposition of presence to absence, of internal and external, of fundamental equivalence and fundamental difference. Librarians have traditionally responded to these complexities by employing a diverse collection of strategies, based on different principles of warrant, ranging from the literary warrant, through to philosophic warrant and to the principles of cognitive authority.

Furthermore, library systems have historically developed within the context of constant, sometimes acrimonious debate from different stakeholders in the information community. From cataloguers who offer suggestions and comments on problematic subject headings, to users who object to the knowledge organization they encounter in libraries, to librarians who develop special schemes for specific interests and communities, to international bodies of professionals committed to the slow and deliberate process of consensus building, the library community airs its issues of warrant in a combination of speech and writing, in physical and digital form, in a way that invites community participation on many levels. As such, library information practice has evolved a complex discourse around questions of warrant that provide a subtlety and richness to knowledge organization that the Semantic Web has not yet attained. If the Semantic Web is to develop further as a new environment for library services, its systems and agents will need to find new ways of incorporating the satisficing strategies that have developed in the library world to deal with the incongruities of information environments.
References
Abstract
Social semantics is more than just tags or vocabularies. It involves the users who contribute the tags, the perceptions of the world, and intentions that the tags are created for. Whilst social semantics is a valuable, massive data source for developing new knowledge systems or validating existing ones, there are also pitfalls and uncertainties. The epistemological analysis presented in this paper is an attempt to explain the differences and connections between social and controlled semantics from the perspective of knowledge theory. The epistemological connection between social and controlled semantics is particularly important: empirical knowledge can provide data source for testing the rational knowledge and rational knowledge can provide reliability and predictability. Such connection will have significant implications for future research on social and controlled semantics.

Introduction
User-assigned tags for classifying or categorizing digital objects have captured attentions from research communities in the last few years. These tags are often one-word or compound-word keywords, category names, or metadata (Guy & Tonkin, 2006) and have many flaws in the eyes of trained information professionals, such as ambiguity, single word restriction, and no synonym control (Mathes, 2004). Researchers have studied this new phenomenon from perspectives of social network (Liu et al., 2006; Mika, 2005), cooperative or social classification (Mathes, 2004), and social annotations of the Web (Wu et al., 2006). A recurring theme in these studies is socially generated semantics, or simply, social semantics.

The term “social semantics” has been used in Agent Communication Language (ACL) in computer science at least since 2000. In the context of ACL, social semantics refers to social commitments that are the beliefs and intentions during agent communication (Singh, 2000). The social semantics approach in ACL involves a social context “in which the given agents participate and within which they communicate” and “allows metacommitments to capture a variety of social and legal relations” (Singh, 2000, p. 34). The social semantics concept in ACL can be used to explain the semantics generated from popular online tagging activities:

- Agents (individual participants) communicate with their beliefs and intentions
- Such beliefs and intentions are public, i.e., observable, in terms of social commitments
- Each communication implies three aspects of meaning: objective, that the communication is true; subjective, that the communication is sincere; and practical, that the speaker is justified in making the communication (Singh, 2000).

Compared to social semantics, controlled semantics (i.e., thesauri, classification, taxonomies, etc.) has a rigid structure and control of synonymous and broader/narrower relationships. Shirky (2005) regards controlled semantics as a way to try to read users’ mind. The “mind reading” nature of controlled semantics, however, can not scale in online social networks. His reasoning is that many participants in online communities are uncoordinated, amateur users and naïve catalogers and there is no authority among them; the large corpus of tags without formal categories, unstable and unrestricted entities, and lack of clear edges all make it difficult for ontological classification to function effect-
ively (Shirky, 2005). Users’ choices of tags represent their beliefs and intentions and they should be the ones who decide the meaning in these tags rather than be forced onto the structured hierarchical categorization.

Many librarians and information professionals do not hold the same position as Shirky does regarding the liberal treatment of folksonomies. Research found that some folksonomy flaws, including misspelling, incorrect coding, and too personal words that would become meaningless without context, were so ill-constructed and contained so many errors and ambiguities that they could derail the value of the emergent social semantics. The term “tag literacy” was even proposed to refer to the “‘etiquette’ of generating tags in a way that increases their social value, balancing individual needs with the needs of the group” (Mejias, 2005).

The varying attitudes toward controlled and social semantics reflect some deeper beliefs and perceptions of knowledge organization among information professionals and researchers. All these raise some fundamental questions: What makes social semantics different from controlled semantics? Where does the common ground lie for the two types of semantics to benefit one another? What implications are there for controlled and social semantics research? This paper analyzes controlled and social semantics from an epistemological perspective and addresses these questions by using examples from both approaches.

**Empirical Representation Versus Rational Organization**

The most prominent attribute of social semantics is the mass participation in describing information objects that used to be the job of trained professionals. Such a description is solely based on their experience and perceptions of the world, which manifests what empiricism in the theory of knowledge defines: our knowledge about a particular subject is “*a posteriori*, dependent upon sense experience” (Markie, 2004). Using an example tag “wedding,” the social tagging results in Figure 1 (a) and (b) present a similar semantic pattern: the tag wedding is related to the things and activities that people use and do in the event.

![Figure 1. Examples of social semantics: Representations for weddings](a) Tags from Del.icio.us (b) Tags from Flickr

<table>
<thead>
<tr>
<th>related tags</th>
<th>(b) Tags from Flickr</th>
</tr>
</thead>
<tbody>
<tr>
<td>flowers</td>
<td>bride, groom, love, marriage, cake, portrait, bride, couple, happy woman</td>
</tr>
<tr>
<td>honeymoon</td>
<td>flowers, bouquet, roses, tag, plan</td>
</tr>
<tr>
<td>jewelry</td>
<td>cake, reception, wedding cake, dance, dancing, food</td>
</tr>
<tr>
<td>places</td>
<td>friends, family, party</td>
</tr>
<tr>
<td>our ceremony</td>
<td>See more in this cluster...</td>
</tr>
<tr>
<td>wedding planner</td>
<td>See more in this cluster...</td>
</tr>
<tr>
<td>reception</td>
<td>See more in this cluster...</td>
</tr>
<tr>
<td>bio</td>
<td>See more in this cluster...</td>
</tr>
<tr>
<td>caribbean</td>
<td>See more in this cluster...</td>
</tr>
<tr>
<td>travel</td>
<td>See more in this cluster...</td>
</tr>
<tr>
<td>vacation</td>
<td>See more in this cluster...</td>
</tr>
</tbody>
</table>
Figure 2. Example of controlled semantics: terms related to weddings in *Art & Architecture Thesaurus*

- **weddings** (rites of passage, <cultural ceremonies>, ... Events)

Terms:
- **weddings** [preferred, C,U,D,American English-P]
- **wedding** [C,U,AD,American English]

Facet/Hierarchy Code: KJKM

Hierarchical Position:
- Activities facet
- ... Events
- ........ events
- ............ ceremonies
- ................ <cultural ceremonies>
- ................... rites of passage
- ....................... weddings

Related concepts:
- distinguished from .... marriage

1. □ ☑ black letter
   (typefaces, type forms, <script and type forms>, ... Components) [3002096x]
   texts, wedding
   wedding typefaces
   wedding types
   types, wedding
   typefaces, wedding
   wedding texts

2. □ ☑ dower chests
   (chests, case furniture, ... Furnishings and Equipment) [300039125]
   wedding chests
   chests, wedding

3. □ ☑ wedding dresses
   (<addresses by function>, <dresses (garments), ... Furnishings and Equipment>)
   wedding gowns
   wedding gowns
   gown, wedding
   gowns, wedding
   dresses, wedding
   wedding dress

4. □ ☑ wedding lamps
   (vertical wine lamps, <lamps by form>, ... Furnishings and Equipment) [3001]
   lamps, wedding
   wedding lamp

5. □ ☑ weddings
   (rites of passage, <cultural ceremonies>, ... Events) [300069158]
   wedding

The tags may seem to lack formality and organization to librarians, but from an epistemological view, these tags reflect a *posteriori* knowledge that the taggers intend to describe. As Liu and Maes point out, social networks (including tags generated by social network participants) “are contextually sophisticated, and framed by our native meta-languages like psychology and perception. Nouns and verb phrases can express ideas and relationships more sophisticated than any RDF or OWL description ever could” (Liu and Maes, 2007, p. ii–iii). As such, social semantics symbolizes people and culture and is specified from bottom up.

Controlled semantics, on contrary, represents a top-down approach in categorizing and organizing information and knowledge. Such a top-down approach follows “certain principles or ideas that form the basis of our understanding of the world” (Southwell, 2007).

Classification schemes, taxonomies, and thesauri as typical controlled semantics carry the attributes of rational organization of knowledge: formal expressions of concepts and intellectual and deductive relationships among the concepts. Terms for the wedding concept in the *Art & Architecture Thesaurus* (Figure 2), for example, are visibly similar to the
social tagging results but exhibit a formality and control. The top figure in Figure 2 shows the hierarchical structure for the term weddings demonstrating to which class the term belongs. In the figure at the bottom part of Figure 2, variant forms of terms are brought together under the preferred term, e.g., “wedding dresses” is the preferred term for wedding gowns, wedding-gowns, etc. It demonstrates the two sources of knowledge — empiricism and rationalism — play different roles in the process of acquiring knowledge.

Connections between Social and Controlled Semantics

Empirical knowledge can be intuitive, demonstrative, and sensitive (Southwell, 2007). Intuitive knowledge refers to something most certain, e.g., black is not white, and demonstrative knowledge involves putting something simple together to form complex ones, e.g., Sun is similar to fire, which is quite certain. The sensitive knowledge is the least certain among the three types because even though something remains logical, it can be false (e.g., all weddings use flowers, rose is a flower, so rose is used in weddings). Social semantics as a form of empirical knowledge allows us to test the truth of rational knowledge, i.e., the logical structures and vocabularies used to reason and organize the empirical knowledge.

The rational nature of controlled semantics provides reliability and certainty. We can expand a concept in information retrieval if we know its broad term or upper class. For example, if we know class A is the parent of class B and class B is the parent of class C, we necessarily know that A includes C. The strength of controlled semantics as rational knowledge lies in reliability and predictability. In an information retrieval system using controlled semantics, a skilled searcher knows when to expand or narrow down the concept of interest. The logical arguments allow the predictable and reliable manipulation of controlled semantics. However, such reliability and predictability is necessarily built upon validity. If the premises (all weddings use flowers, rose is a flower, so rose is used in weddings) are not true, then the argument remains false even though the argument is rational and logical.

The examination of connection between social and controlled semantics brings two points for us. First, since we know if B is a subset of A and C is a subset of B then C is also a subset of A, we can make logical leaps by applying rationalism to experience. In the knowledge organization terms, social semantics represents the empirical knowledge. From a rational perspective, social semantics may not be totally valid because of the reasons discussed in the introduction. However, as empirical knowledge it is possible for us to test it and thus learn from it. We can consider this as the epistemological connection between social and controlled semantics.

Another point is the validity of controlled semantics, which has been rarely discussed in literature. Controlled semantics is constructed by experts. Although studies have repeatedly found the difficulties in using controlled semantics for information searching, the validity question was never raised. The validity issues in controlled semantics include structural and vocabulary appropriateness and timeliness. While evaluating validity of controlled semantics is extremely challenging and sensitive (e.g., how to decide the appropriateness of a term and its place in the structure), lack of empirical data has also made validity evaluation very difficult if not impossible. The new wealth of social semantics would provide the exact empirical source we need to validate the controlled semantics. The emergence of social semantics filled a gap between the rational knowledge and em-
empirical knowledge and for the first time in the history of organizing human knowledge we have the massive empirical data to validate its counterpart—controlled semantics.

**An Epistemological Framework for Social and Controlled Semantics**

The epistemological connection between social and controlled semantics has significant implications for future generation of knowledge systems. The vast social semantic data now available presents new research agenda in designing and implementing new knowledge systems. Research has been well underway to study how the potential of social semantics can be leveraged to develop more intuitive and powerful knowledge systems. Examples include a case study on emergent semantics in the social bookmark site del.icio.us by Michlmayr (2005), extraction of community-based ontologies (Mika, 2005), and feral hypertext in the case of literary and creative hypertext practice (Walker, 2005). Techniques such as ontology matching algorithms, co-occurrences of tags, set theory, and social network analysis techniques have been used to discover semantic relationships between tags and derive ontologies of all kinds (Van Damme et al., 2007; Heymann and Garcia-Molina, 2006). Figure 3 provides a generalization of research in the emergent social semantics. Each type of semantics contributes to knowledge systems by drawing on the benefit from the opposite. Social semantics can provide valuable source for validating and updating controlled semantics while controlled semantics can be used as the source to categorize, cluster, and organize tags. Currently, there are more studies focusing on mining gold from the massive tag repositories than on the other direction.

**Figure 3.** An epistemological framework for social and controlled semantics

<table>
<thead>
<tr>
<th>Social semantics</th>
<th>Knowledge systems that integrate both social and controlled semantics</th>
<th>Controlled semantics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intuitive</td>
<td>Categorizes, clusters, organizes</td>
<td>Logical Deductive</td>
</tr>
<tr>
<td>Demonstrative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitive</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Conclusions**

Social semantics is more than just tags or vocabularies. As stated in the definition of social semantics, it involves the users who contribute the tags, the perceptions of the world, and intentions that the tags are created for. Whilst social semantics is a valuable, massive data source for developing new knowledge systems or validating existing ones, there are also pitfalls and uncertainties. The epistemological analysis presented in this paper is an attempt to explain the differences and connections between social and controlled semantics from the perspective of knowledge theory. The epistemological connection between social and controlled semantics is particularly important: empirical knowledge can provide data source for testing the rational knowledge and rational knowledge can provide reliability and predictability. Such connection will have significant implications for future research on social and controlled semantics.
References
Classification in Scientific Discourse Communities

Abstract
Classifications of natural phenomena demonstrate the applicability of discourse analysis in finding the importance of concepts such as warrant for categorization and classification. Temperature scales provide a body of official literature for close consideration. Official documents of the International Bureau of Weights and Measures (BIPM) reveal the reasoning behind choices affecting these standards. A more cursory scrutiny of the Saffir-Simpson Scale through scholarly publications and documentation from the National Institute of Standards and Technology (NIST) indicates the potential of this form of analysis. The same holds true for an examination of the definition of what is a planet as determined by the International Astronomical Union. As Sayers, Richardson, and Bliss have indicated, there seem to be principles and a reliance on context that bridge the differences between natural and artificial, scientific and bibliographic classifications.

Introduction
Scientific communities have created a wide range of standards that categorize natural phenomena. These categorization standards range from lists to full-fledged classification schemes. Scientific standards for organizing natural phenomena are likely to be based on the need for measurement of empirical variables which is at the heart of the scientific method. Library and information science communities, on the other hand, have developed bibliographic classification to organize knowledge. These two forms of classification as practiced in western culture both stem from Aristotelian logic and practices. Aristotle’s classification in his Generation of Animals not only classified animals, but also applied the conceptual structure of classification that he developed in his Prior Analytics.

Discussion of the differences and similarities between scientific classification and bibliographic classification is sometimes couched in terms of warrant. For example, W.C. Berwick Sayers described what he called natural classification as “made upon the inherent properties in things; upon those properties without which a thing could not be the thing it is.” (Sayers 1926, §36) and what he called artificial classification as based on “some accidental property of the things classified is adopted as the characteristic of arrangement” (Sayers 1926, §35). In his definition, “natural” classification need not be limited to natural phenomena; natural categories can also be effective as a basis for bibliographic classification which implies that “artificial” bases for warrant can be applied to natural phenomena.

Ernest Cushing Richardson made a similar implication when he wrote that “classification is, in its simplest statement, the putting together of like things, or more fully described, it is the arranging of things according to likeness and unlikeness.” (Richardson 1930, 2). He went on to divide “things” into two categories: those outside of “man” — that which makes “man” — that is, nature; and those inside of “man” — that which “man” makes — ideas and art. Richardson clarifies the link between classification of knowledge and classification of nature by pointing out that natural phenomena can be the source of ideas.

Henry Evelyn Bliss went one step further when he wrote that natural classifications, not just artificial classifications, “may have a conceptual and often purposive warp in their fabric.” (Bliss 1929, p. 149). That is, natural classifications may have an agenda; they are not neutral reflections of nature.
To explore the links between natural or scientific classification and classification of knowledge, this study uses discourse analysis of selected standards for natural phenomena addressing the following two research questions:

1. Are scientific categorization standards of natural phenomena subject to the same principles as bibliographic classification (warrant, hierarchical force, etc.)?

2. What discourses operate in scientific communities that shape their categorization standards?

The standards analyzed in this research include various temperature scales developed to measure single variables and two standards used to categorize complex phenomena: those classifying hurricanes and planets. The preliminary findings reported here focus on temperature scales, but include also the bases for analysis of hurricane scales and the definition of celestial bodies. They are examined in relation to bibliographic classificatory discourses. As a discourse analysis, this research is focused on texts and the traces that are found in those texts at close reading. The texts are the data which, when analyzed through close reading, reveal the discourses that structure our understanding. In this instance, I am reading the texts of the standards for categorizing natural phenomena and texts of official bodies that govern those standards through the lense of classificatory principles. This paper reports preliminary results of a project in process, exploring the potential of the texts and the method of analysis as much as any conclusions.

**Temperature Scales**

Temperature scales measure one variable, temperature, for which they require two determining factors: a definition of zero and something to determine the size of a degree. These two factors amount to what we would call warrant. In 1714, Gabriel Fahrenheit set zero as the coldest temperature that could be achieved in a laboratory at the time. He linked the upper point of his scale to the average temperature of the human body (perhaps a slightly feverish one at 100F). By setting zero and 100, Fahrenheit determined the size of a degree. In 1742, Anders Celsius set zero as the freezing point of water and 100 as its boiling point. The difference in these two factors of warrant resulted in two quite different temperature scales as anyone traveling between the United States and the rest of the world can attest. In 1848, Lord Kelvin proposed that zero be absolute zero, the theoretical temperature at which there is no kinetic energy in molecules. He set the freezing point of water at 273.16 K so that the size of degrees would be the same in Kelvin as in Celsius. The less familiar Rankine scale is also measured from absolute zero, but the Rankine degree is defined as equal to one degree Fahrenheit, rather than the degree Celsius as used by the Kelvin scale. The Danish Rømer set zero on his scale using freezing brine. The French Réaumur scale set the freezing point of water as zero and the boiling point at eighty. Others have similar variations. The details are related to the point that is designated zero and the size of the degrees (for example, Celsius degrees are approximately twice the size of degrees Fahrenheit). So the single variable, temperature, is based on two choices of warrant: the source of the two factors. In each instance, some “neutral” scientific basis or warrant was sought.

Factors found in nature have been used as these neutral anchors or set points for temperature scales. If some “natural” set points were sought, it is no wonder that Fahrenheit’s scale, linked as it was to the capacity of early eighteenth-century laboratories, has not satisfied scientists. The discourses that have shaped the modern scientific scales, Celsius and Kelvin, can be traced in the work of the International Bureau of Weights and Measure
(Bureau International des Poids et Mesures, BIPM). The mandate of the BIPM is to provide the basis for a single, coherent system of measurements throughout the world.

In 1887, the International Committee for Weights and Measures (CIPM) approved the Celsius scale as the official *practical* measurement of temperature. It was confirmed by the 1st General Conference on Weights and Measures (Conférence Générale des Poids et Mesures, CGPM made up of government representatives and observers) in 1889 in Paris, thus receiving the imprimatur of the official international scientific community. However, it is important in this discourse to note the designation of its being “practical.”

To anchor this scale to measurable points, scientists at various national laboratories spent the 1890s to the 1920s testing the freezing and boiling points of various elements including sulphur, mercury, zinc, antimony, silver, and gold. They were trying to find reliable ways to define specific points along the temperature scale — not just zero and one other point. They were seeking greater accuracy. Additionally, from the 1890s to the mid-twentieth century, instrumentation for measuring temperature came under scrutiny. Thermometers using mercury, platinum, and gas in particular were tested as another approach to accuracy.

With each CGPM there seemed to be significant discussions. At the 5th CGPM in 1913, national laboratories in Britain, Germany, and the US were encouraged to contribute to defining the standard scale. The 7th CGPM established the International Temperature Scale of 1927 which used the results of the study of different elements’ freezing points to extend the scale to colder temperatures. In the progress of the CGPMs, a significant point came with the 10th in 1954 which adopted Kelvin’s 1848 absolute zero scale as standard for *thermodynamic* temperature. This move was a recognition that context affected measurement. A practical scale suitable for measuring the temperature outside of our windows to determine whether or not to wear a coat is not necessarily the same scale as is appropriate for scientific measurement. We hope not to need to measure the boiling point of gold or absolute zero in our everyday lives, but scientists may need to measure the variable of temperature in such a wide range. So, two perspectives on temperature scales were acknowledged officially at the 1954 CGPM. In 1958, a standard atmospheric pressure was established for the test points that defined the thermodynamic scale to control another variable that affects the measurement of temperature.

With the Celsius scale established as the practical one and Kelvin as the thermodynamic scale, it is not surprising that a quest to bring them together developed. The International Practical Temperature Scale of 1968 established greater standardization in the unit of measurement — the Kelvin — being the same for both thermodynamic and practical scales along with agreement on how specific temperatures are defined (e.g. freezing points of silver & gold) for both. The International Temperature Scale of 1990 continued this constant refinement. The triple point of water (the temperature and pressure at which three phases (gas, liquid, and solid) of water may coexist — 273.16K and 0.01C) rather than boiling point became a set point for both Kelvin and Celsius. The means of defining temperatures continue to be refined. The 1990 scale expands the applicability of gas and platinum thermometers and different elements are used as reference points in different ranges of the scale. This process of refinement continues.

The two discourses of “practical” and “thermodynamic” continue as the BIPM maintains the two temperature scale of Celsius and Kelvin. The refinement of the thermodynamic scale continues with the practical scale tagging along because the two are linked. Who, then, exercises these discourses and why is the thermodynamic scale privileged?
Celsius and Kelvin are both governed by the BIPM. It, in turn, functions under the guidance of the Congress (CGPM) — a sort of United Nations of measurement. The CGPM agenda is supplied from the more select CIPM (International Committee for Weights and Measures — 18 members of CGPM), which is, in turn, supplied by its committees. The Consultative Committee for Thermometry (CCT) is chaired by member of CIPM and the rest of its members are representatives of institutions, mainly national institutes of metrology. So the CCT sets the agenda as long as things run smoothly.

“The CCT was set up in 1937. Present activities concern matters related to the establishment and realization of the International Temperature Scale of 1990 (ITS-90) and thermodynamic temperature, extension and improvement of the ITS-90, secondary reference points, and international reference tables for thermocouples and resistance thermometers.” (CCT http://www.bipm.org/en/committees/cc/cct/) So the agenda remains the same. The majority of the CCT’s energy goes into refining, extending, and measuring the thermodynamic scale. The national laboratories and institutes are understandably more interested in scientific applications than they are in selection of appropriate apparel or when my crocuses will emerge. However, of possible future interest may be the relevance of this refinement in measuring and predicting global warming relative to which even small changes in temperature can be significant. The mundane, readily measurable mid-range of the temperature scales could take precedence. We may see a resurgence of the importance of practical temperature scales and a change in the warrant that defines temperature scales.

Saffir-Simpson Hurricane Scale
A more sophisticated classification is the Saffir-Simpson Scale of hurricanes (the familiar categories one to five) which begins with a complex phenomenon, the hurricane. Saffir-Simpson is based on two variables: sustained winds and storm surges. Wind speed is the deciding factor in categorizing a hurricane — one might call it wind warrant. The Scale is apparently good for predicting what warnings need to be posted. Nevertheless, winds can vary within a hurricane and storm surges are influenced by the shape of the continental shelf and other variables also arise in assessing hurricanes. The Saffir-Simpson Scale is apparently not effective in assessing potential damage. Yet, insurance coverage is triggered by the Saffir-System values assigned to a particular storm and the area affected (Bradley, Hein, and Kruse 2006). Storm surge, the less favored variable, and other factors such as the volume of rain and inland flooding that are influenced by variables such as the shape of the continental shelf and the speed with which the storm moves through an area have a greater effect on damage (slower storms tend to be more destructive). Preventive measures such as levees are also based on what they can withstand on the Saffir-Simpson scale, yet their purpose is to prevent damage. Internationally there are other scales for hurricanes and comparable storms which vary from the US Saffir-Simpson Scale in how they measure these multi-dimensional storms. Wind gusts may be measured instead of sustained winds and even the definition of “sustained” varies (Saffir-Simpson considers “sustained” to be one minute whereas other scales demand five minutes). What discourses of logic, empirical data, economics, and politics determine the scale?

Robert Young, David Bush, and other researchers have been presenting work on the necessity of changing the hurricane classification to the Geological Society of America and other venues (e.g. Young et al. 1999) several times since beginning in 1999 and Jason Carl Senkbeil and Scott Christopher Sheridan proposed a revision of Saffir-Simpson in a
scholarly journal in 2006. However, not until that same year did the National Institute of Standards and Technology (NIST), which governs the Saffir-Simpson Scale, declare its intention to consider the scale’s revision. At that point reconsideration came as a result of findings from the analysis of hurricanes Katrina and Rita. In a press release from June 2006, NIST presented the recommendation that the Saffir-Simpson Hurricane Scale be reviewed. The press release describes the 23 recommendations in a report from an external committee of 26 experts from private, academic, and government institutions (Visiting Committee 2006). A closer examination of these documents, related texts, and perhaps texts from the National Oceanic and Atmospheric Agency is a next step in this study (NOAA does not use the Saffir-Simpson Scale). An close reading of the documentation may be able to address questions such as: Did it really take Katrina and Rita to prompt the review of the Saffir-Simpson Scale? Or, put in knowledge organization terms, did it take a disaster to prompt consideration of what constitutes warrant in categorizing hurricanes? Suddenly and surprisingly, warrant becomes a life-and-death matter.

**Planetary Classification**

A yet more complex task, though without mortal consequences, is the categorization of celestial bodies — notably, the debate that arose over how to define what is a planet. In knowledge organization terms this is a question of concept analysis. In fact, it comes down to the question of the definition of one concept — “planet” — but it has ramifications for other possible categories. Some twenty years ago this categorization was not an issue. There were nine planets and those planets were planets because textbooks and 3-D models showed that they were planets. There were four terrestrial planets (Mercury, Mars, Venus, and Earth), four gas giants (Saturn, Jupiter, Neptune, and Uranus), and Pluto. More recently, however, astronomers have determined that some of the planets thought to have a gas core actually have an ice core. So some gas giants are really ice giants. Pluto turns out to be an object in the Kuiper Belt, but other such objects recently discovered are similar to (such as Santa and Easterbunny) or even larger than Pluto (such as Eris, formerly known as Xena). There are, of course, planets of a comparable nature to all of these types in other solar systems, but in addition astronomers have identified “unbound planets” — planets that have escaped the gravity of their stars. So “planet” is a conflicted category.

This problematizing of the category “planet” highlights the complex efforts to define requirements for what will be considered to be within the category. What will set the boundary? A wide range of bases for warrant is presented in the astronomical literature. That literature recently collided with the popular press as the International Astronomical Union tried to determine whether or not Pluto is a planet. S. Alan Stern and Harold F. Levison (2000) had suggested a variety of requirements for determining the definition of a planet. Two that relate to classificatory factors are that such a definition must: 1) uniquely classify any given body — in classificatory terms that is, it must provide mutually exclusive categories; and 2) be deterministic — that is, the definition must exhibit stability. The discourses in scholarly publications such as those by Stern and Levison were paralleled in the public media and in official press releases of the International Astronomical Union. In the *New Yorker*, reporter Alec Wilkinson interviewed astrophysicist Mike Brown, who discovered Xena, the Kuiper Belt object larger than Pluto. Various discourses are revealed in this popular press article. Brown is represented as having: “a round face and reddish-blond hair, and wears glasses with delicate metal frames in the shape of ovals. He is mar-
ried to a woman named Diane Binney, a school fundraiser, and has an infant daughter, Lilah.” This very human scientist considers (and rejects) giving up “the hard-nosed science view of what a planet is in lieu of a cultural view, and that view includes Pluto, so it includes anything bigger than Pluto.” However, that view adds substantially to the list of planets. Another more scientific take on this option is to define planets as being massive enough that their gravity pulls them together into a roughly round shape. This criterion would allow 23 bodies to be admitted to planethood. Such radical results complicate the warrant for determining the essence of a scientific manifestation, yet they are empirically sound. The upshot is that Pluto is demoted from the original nine, but is the prototype of a new category called “dwarf planets” along with Eris and Ceres??

The International Astronomical Union, as the final arbiter of such matters, proposed a pragmatic solution the development of which is traced through a series of press releases. In this official international body the conflicted scholarly and lay discourses acquired political overtones. The academic (including not only refereed articles, but also the websites of scientists such as Brown) and popular texts joined by the official texts provide fertile ground for unearthing the rhizomatic discourses that construct the current notion of the category: “planet.” The mix of popular and scientific/official discourses well illustrated by a National Geographic contest seeking a new mnemonic clause to help remember the now eleven officially identified planets. Not only Pluto and Eris, but also Ceres, the largest object in the asteroid belt, are now included as dwarf planets which the contest makes legitimate planets so we do not lose Pluto after all. A Montana fourth-grader won the contest with “My Very Exciting Magic Carpet Just Sailed Under Nine Palace Elephants.” (With 2 more … 2008) So while the definition of a planet may be less critical for everyday life, it is clear that the general public is more invested in the discourses around this classification than in temperatures — Celsius, Kelvin, or Fahrenheit — or the categorization of hurricanes.

Conclusion
These examples of scientific categorization standards exhibit both the complexity and diversity possible in scientific discourse. This very variety exhibits the interpretable nature of nature and the possibilities for a range of standards parallel to those for bibliographic classifications. The preliminary analysis performed to date suggests to this researcher that further study is merited. The diverse discourses exhibited in these scientific schemes suggest that empirical evidence and logic are not the only forces operating in the creation of standards — even scientific standards. Cultural, economic, political, and ideological discourses may influence the warrant of scientific categorization of natural phenomena as much as they have bibliographic classification. Choices regarding warrant and other knowledge organization notions such as conceptual analysis and mutual exclusivity appear in both “natural” and “artificial” classifications. My conclusion is that this line of research is worth pursuing because of its potential to reveal the discourses behind both approaches to classification. Future research will also venture into the realm of the metaphysical to ask the question: Are these discourses parallel to discourses that operate in relation to bibliographic classifications, particularly those discourses stemming from logic?
References


Empirical Observation, Rational Structures, and Pragmatist Aims
Epistemology and Method in Julius Otto Kaiser’s Theory of Systematic Indexing

Abstract
Hjørland’s typology of the epistemological positions underlying methods for designing KO systems recognizes four basic epistemological positions: empiricism, rationalism, historicism, and pragmatism. Application of this typology to close analysis of Julius Otto Kaiser’s theory of systematic indexing shows that his epistemological and methodological positions were hybrid in nature. Kaiser’s epistemology was primarily empiricist and pragmatist in nature, whereas his methodology was pragmatist in aim but rationalist in mechanics. Unexpected synergy between the pragmatist and rationalist elements of Kaiser’s methodology is evidenced by his stated motivations for the admission of polyhierarchy into syndetic structure. The application of Hjørland’s typology to similar analyses of other KO systems may uncover other cases of epistemological-methodological eclecticism and synergy.

Introduction: The Theoretical Background
In recent years, the field of knowledge organization (KO) has undergone an epistemological turn, as researchers have increasingly addressed the issue of what theories of knowledge (should) underpin the practice of classification and indexing (e.g., Hjørland, 1997, 2003; Svenonius, 2004; Zins, 2004). In light of KO’s fundamental concern with categorization, it is only appropriate that one product of this interest in theories of knowledge has been a typology of the different epistemological positions that provide its theoretical underpinning. Elaborated by Birger Hjørland in a series of publications (e.g., Hjørland, 1998, 163; 2003, 105–107; Hjørland & Albrechtsen, 1999, 134), this typology recognizes four basic epistemological orientations: empiricism, which posits that knowledge is generated through the observation of phenomena and constructed through the process of induction; rationalism, which holds that knowledge is attained through rational intuition of pre-existent categories and built up by the process of deduction; historicism, which foregrounds the historical and contextual factors that shape knowledge; and pragmatism, which lays stress on the knower’s goals and values, as well as the consequences of his or her conceptualizations for both him/her and the objects with which (s)he is concerned, as key elements in the constitution of knowledge.1

In presenting this typology, Hjørland (1998, 163) distinguishes between the epistemological positions themselves and the methods of classification (or indexing) that they underwrite. This distinction is important, for it suggests that, on an analytic level, the use of any given method within a given KO system can be dissociated from the epistemological position that provides its theoretical justification—that is to say, the methods used in constructing a KO system need not mirror, in all particulars, the specific epistemological commitments of its creator(s) and/or curator(s). Indeed, as Hjørland and Albrechtsen (1999, 133) have observed, actual systems tend not be tied exclusively to methods deriving from a single epistemological position but often incorporate methodological elements derived from different epistemological theories.

1. These lapidary characterizations of the four epistemological positions only summarize, in most general terms, what are highly complex and widely ramified philosophical orientations: for a fuller tabular overview of the basic principles of these positions, see Hjørland, 1997, 60, 71, 74, & 76.
Given that most KO systems are methodologically hybrid, it is legitimate to ask what role Hjørland’s typology might play within KO research. Hjørland and Albrechtsen (1999, 133) have suggested that it can further two lines of research: “to try to illuminate the methodological and theoretical assumptions behind given systems and point out the weak and strong points of these different methods”. In practice, these two lines of research have not been pursued with equal intensity. Considerable effort has been devoted to the comparison and evaluation of epistemologically-grounded methods: one may cite, for example, Hjørland’s (1998, 164; 2005) discussions of the limitations of empiricist and rationalist methods in KO and his calls for greater deployment of historicist and pragmatist methods. Much less work, however, has been done in using the typology as an analytical tool to examine in detail the theoretical presuppositions of individual KO systems. Studies invoking the typology have been more concerned with presenting individual systems as examples of particular epistemological types than with using the types to analyze them: as a consequence, characterizations of the epistemological traits of specific systems have tended to be cursory and sweepingly general in their scope (e.g., Hjørland, 2003, 105, 107). This tendency toward blanket characterization of individual KO systems as exemplars of types runs the risk of foregrounding only one aspect of their epistemological-methodological apparatus. By the same token, it leaves unfulfilled an important aspect of the typology’s potential for enhancing KO research—its utility as a tool in identifying the different epistemologically-grounded elements within a single system and analyzing how these elements interact within the structure of that system.

In this paper, we present a case study of how Hjørland’s typology can be profitably applied to the detailed analysis of an individual KO system. As the object of our analysis we have chosen Julius Otto Kaiser’s (1911, 1926) method of systematic indexing. Originally conceived as a means for organizing information within in a corporate library setting, Kaiser’s indexing theory is best known today for certain features—namely, its stipulations that all indexing terms be divided into mutually exclusive categories of “concretes” and “processes”, which are then to be synthesized into indexing “statements” according to strict rules of citation order—that prefigure key principles of facet analysis (e.g., Svenonius, 1978; 2000, 173–174). When considered in tandem with Kaiser’s (1911, §§ 16–18, 295, 391) firm belief that the purpose of indexing is to analyze literature down to its component “facts” (i.e., informational units) and rearrange them into subject statements with “almost mathematical exactness”, these features may give the impression that his indexing theory was thoroughly rationalist in nature. However, a careful reading of Kaiser’s discussions of systematic indexing reveals that his theory was much more eclectic than is generally recognized. Indeed, as we shall see, his epistemological perspective and his indexing theory incorporated elements from empiricist and pragmatist, as well as rationalist, epistemological-methodological views. Scrupulous and thorough in his exposition, Kaiser gave explicit, and often detailed, justifications for his methodological positions: thus, his writings serve as an excellent venue for (1) using Hjørland’s typology to distinguish the different epistemological elements underlying the methods of systematic indexing and (2) examining how these different elements reinforced one another within the total economy of Kaiser’s indexing theory.

**Between Empiricism and Pragmatism: Kaiser’s Epistemology**

According to Kaiser (1911, §§ 5, 16, 46, 297), systematic indexing concerns itself with the analysis, distillation, and (re)organization of information that has been encoded in
language and recorded in literature. This recorded information, in turn, reflects human experience of, and thought about, the world, for “[w]hat we record is what we observe, what we reason out” (§ 52 [emphases his]): in other words, it represents knowledge (§ 297). Now as pathways to knowledge, the activities of observation and ratiocination require the existence of entities that are the ultimate objects of knowledge. In Kaiser’s view, these objects are (1) “things in general, real or imaginary,” and (2) “the conditions attaching to them”; he calls the former “concretes” and the latter, “processes” (§ 52 [emphases his]). Kaiser’s choice of the term “concretes” as a term for designating things in general is revealing, for it indicates that, despite his stipulation that things can be either “real” or “imaginary”, the prototypical thing is a material object perceptible to the senses (cf. Svenonius 1978, 135): as he put it, concretes “occupy a space, they have a form” and “[e]ach concrete represents something definite to handle” (Kaiser, 1911, § 108). Kaiser’s use of the term “processes” to denote the conditions attendant upon concretes is also significant, for in glossing “processes” as “what [concretes] do or what we can do with them” (§ 55), he attributed a dynamic quality to them. Concretes, then, constitute the underlying substance(s) of the world and processes are their qualities as manifested by their activity: together, they are the sum total of what is knowable about the world (§ 56).

For Kaiser (1911, §§ 53, 56), knowledge is ultimately rooted in the observation of concretes and their conditions. By means of observations of concretes, he argued, we come to form definite conceptions of the entities in the world around us: “[e]ven in their most complex forms—for instance, a battleship specifically pointed out—we know of what they are composed, there is no margin for doubt as to what is included and what is excluded” (§ 108). Yet, Kaiser cautioned, if observation-based conceptions of concretes are definite, they are also strongly circumscribed. This limitedness is due to the fact that “concretes are only known to us superfi cially. … We are unable to give a complete description of any concrete, no matter how many attempt a description” (§ 55)—that is to say, one cannot come to know a concrete by observing it in isolation and intuiting its essence as a Ding an sich. Rather, Kaiser claimed, persons come to know concretes by means of observing the processes in which they are implicated: for example, “[e]lectricity … is a concrete, but it is only known to us by its actions, and it is by observing its actions that we arrive at any appreciation at all of what its probable nature is (§ 55). The upshot of this is that knowledge, for Kaiser, takes on a strongly “phenomenalist” cast: an observer can come to know a concrete only by dint of the qualities that it reveals in its activity within the world.

Besides being dependent on observation of the behavior of concretes in the world, knowledge is also contingent on how this behavior is observed. Kaiser (§§ 55–56) held that “observation is individual” and “the individuality of an observer will be expressed in each of his observations and their application”. He believed that the act of observation is “not an automatic act” but engages the “our mental faculties” (§ 56): it is not exhausted by the merely perceptual action of looking but involves the cognitive activity of seeing. In Kaiser’s view, then, the process of coming to know something through observation is conditioned by the individual observer’s background and capacity for observation: the generation of knowledge is essentially a private, subjective affair.

At fi rst glance, Kaiser’s phenomenalist epistemology appears be quite congruent with certain tenets of classical empiricism. For example, his claim that concretes cannot be known as they are in themselves but only through their dynamic qualities as perceived by the person observing them bears a strong resemblance to the Lockean distinction between
the (unknowable) “real essences” and (knowable) “nominal essences” of things (e.g., Locke, 1975, 417–418). Similarly, his insistence that human knowledge builds on acts of observation that are private and subjective is strongly reminiscent of the “subjective idealism” that characterizes empiricist approaches to epistemology (Hjørland, 1997, 61). There are, however, other aspects of Kaiser’s epistemological perspective that appear to point beyond empiricism tout court.

Standing in tension with the empiricist elements of Kaiser’s epistemology are ideas that are best characterized as pragmatist. Kaiser’s (1911, § 55) view that humans conceive of concretes on the basis of observations of “what they do or what we can do with them” is uncannily reminiscent of Peirce’s (1955, 31) pragmatist maxim, according to which, in observing an object, we must “[c]onsider what effects, that might conceivably have practical bearings, we conceive the object to have. Then, our conception of these effects is the whole of our conception of the object”. In both cases, knowledge of an object is dependent upon observation and analysis of its action in the world. Furthermore, as noted earlier, Kaiser held the epistemological tenet that, since human observers are unable to give a complete description of concretes in themselves and can only form an estimate of their “probable nature” by observing their actions, all knowledge is ultimately provisional in nature:

[E]very new discovery … forces us to modify sometimes some of our fundamental conceptions of concretes, which in turn leads to modifications in our methods of observing and describing them. Hence whatever we assert is always subject to the proviso: at the present stage of our knowledge. (Kaiser, 1911, § 54 [emphases his])

This view of knowledge as provisional and constantly subject to revision is directly comparable to Peirce’s (1955, 4, 54–59) notion of fallibilism, according to which our knowledge of the world is never absolutely certain, but ever revisable in the light of new experience—an idea that historically has been a central Leitmotif of pragmatist philosophy. Moreover, Kaiser’s (1911, § 58) dictum that “individual observation is best followed by individual application of the knowledge gained” points to yet another feature of his approach that is consonant with pragmatism. In his view, the goal of knowledge is not simply to attain a theoretical understanding of the world but to aid in the prosecution of one’s goals: as he puts it, “the philosopher, the scientist and the business man have this in common; each applies individual energy to available information in order to attain the object each has in view” (§ 3). In light of these features, Kaiser’s epistemology seems to lie on the borderlands between empiricism and pragmatism and so reminds us of the sometimes underappreciated affinities between these two epistemological approaches (cf. Kolakowski, 1972, 181–202).

Pragmatic Aims and Rational Structures in Kaiser’s Indexing Method

Kaiser’s inclination toward epistemological pragmatism dovetailed nicely with his broadly pragmatist attitude towards indexing as an activity—an attitude doubtless fostered by his longstanding association with special libraries. In his view, the goal of indexing is to release information from its bibliographical packaging and make it “accessible” and “ready for use” to its users: accordingly, an index “must be so constituted that we can exercise systematic control over the information thus made available” (Kaiser, 1911, § 51, 15). This control of information, however, has to be congruent with the purpose for which it is being organized. In Kaiser’s words, “[a]ll organization has a purpose. The object of organization is not only to control our subjects [i.e., individual information units—TMD]
but to control them so as to give effect to our purpose” (§ 632). Needless to say, this purpose will vary, depending upon the organization for which the index is prepared: different enterprises will have different purposes and the index created for each enterprise will reflect its particular concerns.

The particular purpose of a given index, according to Kaiser, should determine both its organization and contents. In his words, “[o]ur plan of control will … depend upon our purpose” and so “[m]ethods must of necessity always be subordinate to purpose” (Kaiser, 1911, §§ 106, 248)—that is to say, the purpose for which the index is being prepared should govern the process of organizing the information at hand. This information, however, should consist only of that which is considered relevant to the purposes of the enterprise in question:

We want to be informed—and we want all possible information—on that which has a direct bearing on what we are concerned with. Indexing by which we make our information accessible [sic] has therefore a negative and a positive function[:] it throws out what is not required, it concentrates on that which is required (§ 45).

By calibrating both the selection and organization of information to the perceived needs of an enterprise, “we reduce our materials to that which is essential for our purpose, we create a nucleus of effective information, information which will be of real use to us in the pursuit of our business” (§ 46). In general, then, Kaiser’s vision was one of an index “developed on the basis of the analysis of goals, values and consequences”—the very essence of a pragmatist KO system as defined by Hjørland (2003, 105)—within the framework of a corporate library.

Although Kaiser envisioned the function of indexing in pragmatist terms, his indexing method was based largely on rationalist mechanisms. The lexical content of the index had an empirical basis, for Kaiser (1911, § 114, 318) advocated using words and word phrases extracted from the text being indexed as indexing terms. Once a term had been selected, it was subjected to rather stringent regimentation. Each term was assigned to one of a triad of categories representing an expansion of Kaiser’s original dyad of entity types: “concretes”, “countries” (technically, a subclass of concrete), and “processes” (§§ 299–301). These categories were to be mutually exclusive: to this end, Kaiser (1926, 23, § 11; 27–28, §§ 22–25) proposed a series of ways to disambiguate potentially problematic terms (e.g. “organisation” is a concrete term, while “organising” is a process term), going so far as to semantically factor compound words whose components designated concretes and processes (e.g., “bacteriology” became “bacteria” [concrete] + “study of” [process]). These fundamental categories were combined into more complex strings of terms known as “statements”, whose composition was strictly limited to three possible syntactic forms: (1) Concrete–Process, (2) Country–Process, and (3) Concrete–Country–Process (Kaiser, 1911, § 302). An important consequence of this syntactic regimentation was that only concretes or countries could serve as main index entries, while processes occurred solely as subdivisions—a reflection of Kaiser’s (§§ 302, 313, 384) view that, in business contexts, concretes should have primacy over processes. The main term of each statement (i.e., the country or process) was then related to other main terms by means of a syndetic structure recognizing synonymy, hierarchical (i.e., general term–specific term), and, to a lesser degree, associative relationships (§§ 414, 416, 423).

This proto-faceted indexing method was not immune to the limitations that can hedge rigorously rationalist approaches: for example, as Kaiser himself admitted, the categories of “concrete” and “process” could not accommodate certain classes of terms (e.g., math-
The semantic factoring of terms couldn’t account for polysemous compound words; and the strict syntax of subject statements imposed a structural limitation on their expressiveness (Kaiser, 1911, § 327; 1926, 27–28, §§ 23–24). There was, however, one respect in which Kaiser’s pragmatist vision of indexing influenced the articulation of his otherwise rather rigid rationalist structures. This was his admission of polyhierarchy into the syndetic structure of the index. As the following passage shows, Kaiser’s acceptance of polyhierarchy was motivated by a classically pragmatist motive:

Take any commodity: it has a number of properties, and may be viewed from a number of different aspects. Logically it belongs to just as many classes or classifications as it has aspects. … [B]y means of related terms we may … combine them [sci., main entry terms—TMD] in as many logical classifications [sci., taxonomical relationships—TMD] as may be called for by our business interests. We can choose these connections or classifications in strict accordance with our business needs, and we are of course presumed to be expert in our own business. (Kaiser 1926, 26, § 20)

In arguing that a given main entry term could serve as a subclass term to different superordinate class terms and so stand in a number of taxonomical relationships on the basis of the anticipated “business needs” of the users for whom the index had been prepared, Kaiser was able, at least in part, to wed his pragmatist aims with the rationalist mechanisms of his index.

Concluding Remarks
Kaiser believed that the validity of his systematic indexing would ultimately stand or fall with the degree of consistency that an index could achieve in its structure and so sought to create as consistent a rationale for it as possible: in this, he gave voice to a rationalist impulse that characterizes both his system and his mode of exposition. However, absolute consistency is an elusive goal and Kaiser did not entirely succeed in this enterprise, for neither the epistemological bases nor the methodology of his indexing system are free of inconsistencies and paradox (cf. Svenonius, 1978). In using Hjørland’s typology of epistemological positions to an analysis of systematic indexing, we have sought to pierce Kaiser’s discursive veil of consistency, uncover the hybrid nature of his epistemology, and limit the ways in which his epistemological position(s) related to his methodological prescriptions.

As we have seen, Kaiser’s epistemology combined elements of empiricism and pragmatism, whereas his indexing method involved rationalist mechanisms within a broadly pragmatist framework. The prominent role of pragmatist ideas in both Kaiser’s epistemology and his methodology is striking, especially in light of his reliance on rationalist mechanisms for index construction: the intersection of these two approaches in the methodology of systematic indexing suggests that the gulf between them is less extensive than has sometimes been suggested. Also noteworthy is the fact that, despite its eclectic epistemological-methodological bases and some inconsistencies in structure, systematic indexing proved, on the whole, to be a fairly effective indexing method for industrial libraries (cf. Barbour *apud* Kaiser, 1926, 37–41): indeed, its hybrid nature sometimes enriched its structure, as is evidenced, for example, in the interweaving of pragmatic aims and rationalist structures in its polyhierarchically inflected syndetic structure.

In conclusion, we note the utility of Hjørland’s typology in uncovering these otherwise hidden aspects of Kaiser’s system. Its application to the close analysis of other KO systems has the potential to reveal similar epistemological-methodological eclecticism and synergies in systems that, at first blush, seem monolithic in their approach: such analyses may well improve our understanding both of historical KO systems and the various ways
in which different epistemological positions can interact within the limits of a single classification or indexing scheme.

References


Richard P. Smiraglia — Long Island University

Noesis
Perception and Every Day Classification

Abstract
Perception is a crucial element in the viability of any knowledge organization system because it acts as a filter that provides contextual information about phenomena, including potential categorical membership. Perception is moderated culturally, but “social” systems exercise little or no cultural conformity. “Every day classification” is rife throughout human experience; but classification arises as a system of formal constraints that embody cultural assumptions about the categories that are the products of human cognition. Noesis is a perceptual component of Husserl’s phenomenological approach to human experience. How we perceive a thing is filtered by our experiential feelings about it. The purpose of this research is to increase understanding of the role of cognition in every day classification by developing a fuller profile of perception. Photographs of mailboxes (a mundane, every-day example) from different locales are compared to demonstrate the noetic process. Tag clouds are analyzed to demonstrate the kinds of perceptual differences that suggest different user perceptions among those contributing tags.

1. Perception and KOS
This paper is about the role of human perception in every day classification. Systems for knowledge organization (KOS) occupy a broad spectrum—from universal bibliographic classifications to emergent domain-specific taxonomies to social tagging—with many variations between. On many levels it is appropriate to suggest that perception is a crucial element in the viability of any KOS. Related to cognition but colored by individual experience, the perception of individual humans acts as a filter that provides contextual information about any phenomenon, including its potential categorical membership. Perception is moderated culturally—that is, although our perceptions are very much existential, they also are moderated by the cultural forces that provide the context in which we live. Sounds and sights that are considered uplifting in one culture might cause shock or distress in another. Formal systems for KO operate in specific cultural milieu that require the user to conform perception to the reality modeled in the system (consider e.g., Medical Subject Headings (MeSH) or George Sherman Dickinson’s Classification of Musical Compositions, to name two). But recently emerging so-called “social” systems exercise little or no (or at least very subliminal) cultural conformity. Every tag cloud is dominated by terms that can be seen clearly to conform to some norm among the taggers, but that core is surrounded by variability—tags that appear semantically to be far removed from the core. Perception is obviously one epistemological element influencing that variability. The purpose of this research is to increase understanding of the role of cognition in every day classification by developing a fuller profile of perception.

2. Every day classification
Of course, one can also posit a kind of “every day classification,” which is rife throughout human experience. Every human action involves decision-making, which by its nature produces categorization; everything from the most simplistic (e.g., inside/outside, daytime/nighttime, hot/cold, safe/dangerous, etc.), to the complex (e.g., fruit/nut/meat, or cheap/costly/expensive but worth it, for example) sets up essential classifications of what might otherwise be considered intuitive knowledge. Jacob (2001) reviewed several approaches to understanding the human processes that contribute to a “cognitive core”
(p. 81) and suggested that classification needs to be analyzed “by studying its impact within the settings of everyday activity” (p. 96). In an earlier paper Jacob (1994) emphasized the importance of categories as “building blocks of cognition” (p. 101). She also made an important distinction between the concepts of “categorization” and “classification,” by contrasting the cognitive function of the former with the formal systematization of the latter. That is, as noted above, humans engage in categorization constantly as part of the experience of being. Classification arises as a system of formal constraints that embody cultural assumptions (or demands) about the categories that are the products of human cognition. Jacob suggests rigid schemas must be rejected because of the relative contextual experience of participants. This suggests that traditional, formal approaches to classification might be better accommodated with more fluid structures.

The environment of the semantic web is one in which mutuality is encouraged and a multiplicity of vocabularies reside side-by-side. Jacob’s approach to cross-disciplinary communication suggests one way of looking at the issues of vocabulary regulation and category inflexion in the multiplexity of the semantic web environment. No particular, rigid, language or set of categories is preferred. Rather, many are allowed and crosswalks are created to encourage movement back and forth.

I would like to suggest that in addition to Jacob’s notion of contextualization, we also have to comprehend noesis (perception) when we look at the human (or every day) conceptualization of classification as a social action, in the semantic web or elsewhere. One man’s beer is not necessarily another’s biertje. The notion of “beer” is shaped by each person’s expectations in the moment of encounter. What good does it do to posit a category, if every interpreter’s inflexion of it is different? The answer is visible in the disparity in terminology seen in the tag clouds that emerge as the products of today’s social classification. Munk and Mørk (2007) point to the notion of least effort to explain the bandwagon effect seen in tagging environments. But as Kipp and Campbell (2006) pointed out, variation in terminology is rife. Some of the variation is likely due to differing perceptions of the individuals involved. One good way of understanding the meaning of perceptual variability is noesis.

3. Noesis

Noesis is the perceptual component of Edmund Husserl’s phenomenological approach to human experience. Husserl’s mature thought arose from mathematics and logic, and in the end produced what he called a phenomenological method, which was to be seen as positioned over and against psychologism. Standing alongside, but considerably philosophically different from, logical positivism, Husserl’s phenomenology proceeds from much earlier work by René Descartes. In particular, Husserl proceeds from Descartes’ famous maxim cogito ergo sum (Latin: ‘I think, therefore I am’). For Descartes this was a stepping off point for defence of the role of deduction. For Husserl it is the beginning of perception as an essential component of analysis. In a series of lectures delivered at the Sorbonne in February 1929 (the content of which was later published as his Cartesian Meditations), Husserl developed his notion of transcendental phenomenology. There is not space here to rehearse the entirety of Husserl’s philosophy (and neither is that the point of this paper), but we can introduce the concept of noesis by using an outline of the phenomenological process.

For Husserl, it all begins in a primary instance with the Ego, which is all that is. In the beginning of perception, nothing is, except that which is perceived by the Ego. The
method of perception entails a sequence of *epoche*, brackets around specific entities in the perception of the *Ego*. The *epoche* is the method by which one might apprehend oneself by bracketing oneself over against the contextual world. Any spatiotemporal thing that belongs to the world exists for *Ego* if it is perceived by *Ego* (Husserl 1950 [1999], 21). *Eidetic description* is a process that isolates specific entities for analysis by transferring empirical descriptions into the dimension of perception (Husserl 1950 [1999], 69). Each isolate consists of its experienced form (*cogito*) and its concrete form (*cogitatum*). Perception takes place in a sequence of temporal acts (*cogitationes*). Experience, then, is a matter of synthesis, in fact of the synthesis of syntheses. I see many things at once and it is their contextual synthesis that becomes for me reality.

**Figure 1a. Eidetic description**

![Diagram of Eidetic Description](image1)

At the same time, for each thing that I see, I see in fact, a sequence of things, which must be synthesized.

**Figure 1b. Eidetic description is synthesis**

![Diagram of Eidetic Description as Synthesis](image2)

Each glimpse of the world reveals a collectivity of isolates (*cogitos*) that is perceived as the collectivity of a sequence of glimpses (*cogitationes*), each of which leads to its own *eidetic* process as well. *Noesis* comes into play at the point of analysis at the end of each arrow, or in Husserl’s terms, at each *eidetic* moment. At each analysis, we bracket an isolate to analyze it. The analysis is *noesis* and the analysand is *noema*. *Noesis* is the busying of the ego into whose vision the *noema* enters. Every isolate is perceived (comprehended unconsciously) as an element of a larger scenario, all of which have meaning against the personal experience (*Ego* acts) of the individual who is perceiving.
Consider a concrete example. I am aware that I exist, because I am aware of the functioning of my Ego over against the context of my surroundings (at the moment two computers, some lights, a bookcase, some tulips, and a cup of tea). Let us say I wish to focus on the tulips. To create such a focus is to bracket them (epoche). I see in front of me two yellow flowers (cogitationes) in a glass vase, which I determine to be tulips (cogitos). How do I perceive the tulips? I perceive them through my prior human experience, and as part of a whole milieu. I see them as representative (noema) of life, springtime, sunshine; I see them as colorful against this boring backdrop; I see a reminder of the Netherlands, which is a place of which I am fond and a place in which there are many tulips; I see a clear glass vase with just the right amount of cool water and the erect stems as alive and well nurtured—all of this is noesis. I don’t see flowers, I see the collectivity of my prior experience of them.

What does this tell us about everyday classification? It says that it is too simplistic to generate a category of “tulips” from mere cognitive function, without also engaging the fluidity of perception that might reveal cross-noetic dialogue. Noesis tells me many things about the tulips and requires me to synthesize them, not just by simple addition into a set, but also by prioritization. The experiential feelings are acts of my Ego, perceiving and synthesizing against an eidetic backdrop of context. What this suggests is not the usual reflection that the presence of a thesaurus would “help” (see Noruzi 2007), but rather, that multiplicity, such as that suggested by Jacob, should not only be encouraged but also facilitated. Noesis suggests the potential richness of everyday classification.

4. Mailboxes — *Noesis and Cultural Moderation*

In order to extend the noetic metaphor and simultaneously to embrace the concept of cultural moderation (through either variation or conformity) we use the mundane example of a postal box (or mailbox). Photographs of mailboxes from different locales are brought together for comparison (*eidetic reduction*, if you will). These appear in Figure 2 below.

*Figure 2. Mailbox noesis*
By positing the *eideia* (here the “mailbox”) we “bracket” the phenomenon and instead focus on individual perception and varying realizations. We see that mailboxes are devices for communication (for posting letters, but more directly as loci for handbills and graffiti); we see that as icons of government mailboxes are variously efficient, sleek, sturdy, decorous, pious, modern, and timeless. This is a very simple illustration but it already demonstrates the potential power of a *noetic* analysis of every day classification. And we also see that the simple concept of “box for the mail” comes out differently in each culture. It is a simple but eloquent example of *noesis*—the power of perception in a cultural milieu to shape even the contents of a mundane class of everyday objects. Combined with the *noetic* process we can see that despite the sameness of function, the influence of cultural priorities dictates differences in expression or realization, which in turn affect the *noetic* character of the *eideia*.

So, there are at least two ways in which *noesis* can be useful in the analysis of this single phenomenon. We can bracket the idea of “mailbox” in order to see how the idea is realized differently, especially as a result of cultural moderation. But we also can use the literal image of one of these mailboxes to walk through the phenomenological method, to reveal the *Ego* acts involved in *eidectic reduction* as the eye surveys the landscape, discovers the mailbox in context, and deals with the phenomenon from different perceptual reflexes.

**Figure 3. Noetic synthesis**

In Figure 3 we see this mailbox in the distance, at the end of a sidewalk, a canal sits behind it and across the canal are some residential buildings. As we approach the mailbox it becomes more pronounced in our perception, the buildings across the canal grow more prominent as the trees recede. The mailbox itself, at last immediately in front of us, occupies our entire spectrum, but as we walk behind it we see it recede against the backdrop of a large school building, and then we see it adjacent to a bicycle rack. The mailbox itself changes only slightly as the *noetic* process proceeds, and as slightly different perspectives reveal the entirety of the shape of the box. But *noetically* we see that the mailbox has a central role in the community life of this city (Amsterdam, obviously), standing here adjacent to residences, a school, along a street, a canal and at the intersection of sidewalks, and alongside a bicycle parking spot. The mailbox, then, becomes a part of the civilization in which we find it, and it also is reinforced as a communicative tool helping to link citizens. This communicative perception is doubly reinforced by the mailbox’s proximity to modes of transportation. Thus our synthesis reveals not simply the isolate (*cogito*), but also the cultural moderation that surrounds it and from which it springs. Like our tulips above, the infixion is both complex and fluid. This expanded perception is one of the advantages of *noetic* analysis.

5. *Noesis* and social tagging

*Noesis* helps explain perceptual differences that arise in even social tagging environments. Kipp (2007) has noticed affective use of tags, which appear to be contextually irrelevant to the casual observer but seem to be meaningful to taggers. She suggests that
taggers see their relationships to documents in many different ways; in particular some seem to express emotional and personal relationships. There is a sort of foggy cloud surrounding the core concepts in any tag-cloud, and in the fog lie the individual perceptions of specific taggers. For those who see a mailbox as benign the core tags will do; for those who see it as an instrument of bad news or good news, different noetic values will attach themselves to their tags. Perceptual differences are a sign of noetic activity in this very prevalent form of every day social classification. Several tag clouds (formed from searching “postal” and “mailbox”) from Del.icio.us are reproduced in Figure 4, to demonstrate further.

**Figure 4. Postal tag clouds**

These tag clouds are all associated with postal services, and we see the conforming core terms associated with postal functions. But we also see terms that evoke noesis—in the “postal” group for instance, we see terms associated with government, shopping, productivity, pranks (including strange, random, and weird) and with surveillance. In the “mailbox” group we see terms such as lifehacks, mashups, cool, and bizarre—terms that at first glance seem far removed from the reality of a mailbox, but that reveal cultural contexts that are infixed with mailboxes in the perceptions of various taggers.

### 6. Conclusion

There is no doubt about the role of cultural conformity in everyday classification. Every act of perception, as every act of classification (including tagging) is moderated culturally. Even the simplest isolate is infixed in a noetic synthesis that carries potentially burgeoning layers of meaning. Fortunately, the fluidity of categories that is required for truly communicative, cross-domain everyday classification, is revealed in the tag clouds when
we apply the concept of *noesis* to analysis of the clouds. A fuller profile of perception reveals the role of cognition in everyday classification as well. Whereas formally structure classes involve the use of strict “IsA” rules, *noesis* demonstrates the fuzzy boundaries that humans carry with them in every *ego* act. One of the advantages then, of what seems to be uncontrolled social classification, is the ability of the tag cloud to reveal the *noetic* infixions that surround any given isolate, even as the user group’s cognition shifts with ongoing usage.

*Noesis* then, describes an epistemological foundation that is essential for post-modern knowledge organization systems. Social contexts, cultural moderation, and perceptual fluidity are constants in the *ego* acts of classification. A fuller profile of perception, which might begin with further exploration of the power of *noetic* analysis, is required for the development of 21st century knowledge organization systems.

**Acknowledgement**

I would like to thank David Jank for his advice in the construction of this narrative.

**References**


——. 2007b. E-mail 15 Nov. 2007.


Munk, Timme Bisgaard and Mørk, Kristian. 2007. Folksonomy, the power law & the significance of the least effort. *Knowledge organization* 34: 16–33.

Deliberate Bias in Knowledge Organization?

Abstract
“Bias” is normally understood as a negatively loaded word, as something to be avoided or minimized, for example, in statistics or in knowledge organization. Recently Melanie Feinberg suggested, however, that “if we cannot eliminate bias, then we should instead attempt to be more responsible about it and explicitly decide on and defend the perspectives represented in information systems”. This view is linked to related views: That knowledge organization is too much concerned with information retrieval and too much described in the mode of scientific discovery, as opposed to the mode of artifact design: “From the literary warrant of Hulme to the terminological warrant of the Classification Research Group (CRG), to Hjørland’s domain analysis, the classificationist seems like one who documents and compiles, and not one who actively shapes design.”

This paper examines these claims, which may be understood as questions about subjectivity and objectivity in classification and about positivism versus pragmatism in research. Is KO an objective and neutral activity? Can it be? Should it be? A dominant view has been that knowledge and KO should be understood as a passive reflection of an external order. This has been termed the mirror metaphor of knowledge and is related to empiricism and positivism. The opposite view — which is in accordance with both Feinberg and Hjørland — states that knowledge organization should be functional and thus reflecting given goals, purposes and values. It is related to pragmatism in philosophy.

Introduction
Melanie Feinberg has recently suggested some important principles in the theory of knowledge organization (KO). Her main suggestions are 1) That knowledge organization should be seen as a process of active construction rather than as passive discovery (Feinberg 2007a) 2) that a knowledge domain is not a single entity identified by expert consensus, but that information system designers should describe and defend the criteria by which they construct a domain. Her prize-winning paper (2007b) contains the following conclusion:

I have argued that Hjørland’s description of domain analysis seems to imply the concept of a single domain per subject area, structured by a common set of principles (specifically epistemological underpinnings), with a reliance on expert opinion to evaluate the stability of the approaches are seen as undergirding each domain. This single-domain approach thus appears to rely on assumptions that may not hold true in all situations and that constitute a form of unacknowledged bias. A single domain necessitates ill-defined domain boundaries (which presents additional problems due to multidisciplinarity) and is inevitably incomplete, while lacking the requirement to explicitly justify what the domain includes and how it is structured. In contrast, recognition of multiple domains for a single subject area requires that an information system designer describe and defend the criteria by which a domain is constructed. Her prize-winning paper (2007b) contains the following conclusion:

The present paper contains arguments that the views of Hjørland and Feinberg are basically in agreement rather than in conflict with each other as suggested by Feinberg. Furthermore it seeks to clarify some issues in the domain analytic position.

Describing and classifying domains of knowledge
What is psychology? What is archeology? What is a domain? It is hard to deny, for example, that Sigmund Freud and B. F. Skinner are famous names in the history of psychol-
ogy. However, their understanding of what psychology is was very different. One could possibly argue that their conception of psychology was so different that they excluded each other from the domain: Skinner accepted only observable stimuli and reactions and their precise measurements as belonging to the field of psychology, whereas Freud saw such studies as providing trivial information with no real importance, for which reason he constructed psychoanalysis as a separate field. Yet we consider them both parts of the history of psychology. Gjerløff (1999) writes that before 1830 history and archaeology were considered one discipline, but between 1830 and 1890 an understanding was developed which separated them. This self-understanding is connected with a certain view of what historical sources and archaeological sources are and how they are interpreted. Gjerløff argues that this understanding is problematic and the two fields need to be better integrated and have a broader understanding of the concept of sources. It may thus be argued that the understanding of archeology as a discipline separated from history is just a historical construction which has served some interests, perhaps first and foremost some professional interests among professors and museum directors. We may agree or disagree with Gjerløff’s point of view. If we agree, it has important consequences for how knowledge in the fields of archeology and history should be organized. On the other hand, if one librarian agrees with Gjerløff and the rest of society still finds that those two fields are distinct and different, should this librarian just disregard all other opinions? Is description and classification just based on individual opinion? Is any opinion as good as any other? If this last question is the case, then any kind of KO is as good as any other, implying that the field of KO is without a methodological basis.

It is important to realize that disciplines are dynamic and theory dependent. They are also dependent on social power structures, first and foremost in the form of financial support. Toulmin (1972) differentiates between the content-knowledge of a science and the institutional aspects of science, such as the professional forums. He suggests that science is generally continuous because either the content or the institution will remain stable while the other changes. In response then the first will adapt, in an iterative process of constant change and constant stability. There is continuity because each generation is always taught by the preceding generation of scientists and also because the research questions in which a community is interested are predicated on the current concepts they hold, even when the results of such research might indicate that changes are needed to better adapt the concepts in response to other concepts or other facts about nature. Toulmin’s differentiation between content-knowledge and institutional aspects corresponds to Hjørland & Hartel’s (2003) ontological versus social dimension of a domain.

To describe the history of a field is a creative and constructive process. Consider the field of Knowledge Organization. On the one hand we have a field which is in some ways “given”. We have, for example, this ISKO conference. In other ways, however, each of us construes the field. Some among us work from the perspective of knowledge management, some from the perspective of facet-analysis, some from the perspective of bibliometrics and so on. Each course on KO, each textbook, each history of the field and each classification scheme describes and classifies the field according to the perspective of the author or designer. Feinberg is certainly right: Knowledge organization should be seen as a process of active construction rather than as passive discovery. Anybody concerned with the field of knowledge organization should describe and defend the criteria by which they construe this domain. Researchers in KO are not working on a different project compared to the people doing indexing, classification and organization of the literature of
the domain. Inside the domain, however, many different perspectives compete with each other. Each researcher as well as classificationist will always be biased and subjective, e.g. towards a bibliometric or facet-analytic understanding. This leads us to the question of the role of subject specialists in KO.

**The Role of experts in Knowledge Organization**

Subject knowledge is required in order to do a qualified job in KO. People who are not familiar with discourses in the domain to be organized are not equally qualified to construe the field. Above we considered Knowledge Organization as an example of a domain. Any researcher in this field has a perspective, which may be more or less in agreement with the perspective of other researchers. As indicated above many different perspectives seem to compete today in this domain. These competing perspectives may be differently distributed among, for example, researchers and indexers. However, both groups try to contribute to the fruitful development of the field based on their knowledge and job tasks. To make choices among different perspectives presupposes knowledge about these perspectives. This point seems to be in conflict with what Feinberg finds:

This identification of experts being closer to “truth” than laypeople seems also to imply that, even in cases of debate among experts, the dominant or majority view is typically the one that “follows the general rule” and thus is more “valid.” (Feinberg 2007b)

The view that knowledge organization should be based on the consensus or dominant view among experts was expressed by Henry Bliss (1929), it is not a view expressed by Hjørland anywhere. On the opposite, for example, Hjørland (1997, p. 172) quoted Søren Kierkegaard:

> The truth is always in the minority; and the minority is always stronger than the majority, because the minority is ordinarily composed of those who do actually have an opinion, whereas the strength of the majority is illusory, composed of the crowd which has no opinion—and which therefore the next minute (when it becomes apparent that the minority was the stronger) embraces the opinion of the minority, which now becomes the majority, that is, the opinion becomes rubbish by having statistics and the whole crowd on its side, while truth is again a new minority.

As far as truth is concerned, the same thing happens to this awkward monster, the majority, the public etc. as we say happens to the person traveling for his health: he always arrives one station too late. (*The Journals of Søren Kierkegaard* X(3) A 652, in 1850)

Feinberg’s attribution of the majority view among experts to the domain analytic theory of Hjørland is thus simply mistaken. Hjørland is therefore much closer to Feinberg’s own position than what appears from her papers: Hjørland also encourages information specialists to form their own opinion and to contribute activity in the design of knowledge organizing systems. The most important difference seems to be that Feinberg assumes that this can be done without participating in or relating to the discourses in the domain, without proper subject knowledge. Feinberg draws on the feminist scholar Donna Haraway to be discussed below. However Haraway as well as Hjørland do not put themselves outside scientific and scholarly research. It is important to realize that contributions to the production, use and organization of knowledge cannot be made from neutral positions, outside the struggle within domains. The organization of knowledge is part of the struggle within fields to contribute to the development of those fields. There has been and still is a strong ideological tendency within librarianship to disengage the field from discourses in the fields being organized. This (positivist) tendency may be the most important blocking for the advancement of LIS as a scholarly field.
The point of view of Domain Analysis is that subject knowledge is necessary, but not sufficient. Information specialists need other kinds of knowledge about a domain compared to what ordinary subject specialists typically do. This is further described in Hjørland (2002). Ørom (2003) is a contribution to a domain analysis of arts studies. Librarians and information specialists who have read and understood this paper should clearly be better qualified to organize and retrieve knowledge in this domain compared to people which do not possess this knowledge.

**The “family” of pragmatic philosophies**

Feinberg bases her view on the feminist epistemology of Donna Haraway (1988), but remains critical to Hjørland’s “realism”. This is strange since the last is developed on the same kinds of thinking as that of Haraway (and Sandra Harding, among others). These feminist epistemologists are “realists” in the same sense as Hjørland: “a no-nonsense commitment to faithful accounts of a “real” world” (Haraway 1998, p. 579). According to Code (1998) the main contribution of feminism to epistemology has been to move the question ‘Whose knowledge are we talking about?’ to a central place in epistemological inquiry. In this respect feminism is related to anthropological studies and critical epistemologies such as critical theory, all of which have demonstrated the limitations in traditional epistemologies as depending on the world-view of white, Western men from upper social classes. The implication of this view is that feminist epistemologists are producing conceptions of knowledge that are specifically contextualized and situated and of socially responsible epistemic agency and thus related to hermeneutic and pragmatic philosophies. Many feminists retain a realist commitment to empirical evidence, while denying that facts or experiences can be taken at face value. The term “standpoint epistemology” is used by some feminists to underline that an epistemological position cannot be neutral, and therefore simple honesty demands that researchers should explicate the goals and values that guide their investigations.

Harding attempts to reconstruct the veritable norm of ‘objectivity.’ She seeks to replace the ‘weak objectivity’ of the male-dominated scientific world — a pseudo-objectivity riddled with value-laden theories, political biases, domineering interests, commodified research, and blinkered ethical vision — with the ‘strong objectivity’ that comes only from a ‘robust reflexivity’ attained through a rigorous self-scrutiny of one’s socio-epistemological starting point. Harding notes that the very concept of ‘value-free knowledge’ is oxymoronic since the goal of being disinterested is an interest in itself, and it allows science to separate fact from value and abrogate responsibility for its actions. Since ‘value-free’ theories are impossible, Harding argues, one might as well acknowledge the values that inform one’s research, be it to make money or to improve the lives of the sick, and debate their comparative validity, and struggle to have science informed by progressive interests. (Best 1998)

Olson (2002) from a feminist position realized that “The names we give things colour the ways we perceive them. Those in a position to name hold the power to construct others’ perceptions and realities”. Recently Furner (2007) suggested “A Critical Race-Theoretic Perspective” in KO. This may be considered part of the same “family” of epistemologies as the pragmatic, critical and feminist perspectives. They are all very fruitful. However, there is a need to concentrate on what they have in common, what principles of Knowledge Organization they imply, and which other families of KO they are opposed to. In this way we could possibly avoid making unnecessary divides between us and instead cooperate in improving the knowledge base of our field.

Another important issue is to generalize the epistemological basis of KO. The perspectives presented in this section may seem relevant for a better representation of religions,
sexes, races, cultures and social classes in KO. But are they also relevant for the hard sciences? How should we organize, for example, nations, animals, diseases, chemicals and stars? What kind of epistemological questions are at play in such cases? We have much too little research on such principles. In all these cases there is a fundamental conflict between “positivist” and “pragmatic” views: Positivism may be understood as the view that knowledge and KO should be understood as a passive reflection of an external order. This has been termed the mirror metaphor of knowledge. The opposite view is that knowledge organization should be functional and thus reflecting given goals, purposes and values. But is it really possible to argue that scientific phenomena should be classified according to human interests rather than objective structures discovered by science? We will now turn to this question.

Degrees of freedom in organizing knowledge

In the Danish classification system “DK5” the Virgin Islands are classified as part of “Denmark” (46.5) even though they were sold to the USA in 1916! This can be seen as a “bias”, but it may be a well considered bias in that the users of the system will primarily be interested in these islands because of their former relations to Denmark (Danish books are written on these former Danish possession for a Danish audience). According to pragmatic theory, concepts are defined by pragmatic criteria by different kinds of interests. Nobody cared, for example, about how far into the sea a country has jurisdiction until interests such as fishery and potential extraction of resources made this relevant. Any application of concepts and any KO thus have to consider which definitions and semantic relations are purposeful in the given context. This point of view is contrary to the widespread ideal within LIS to standardize meanings.

Information services exist in order to support some kind of goals and values. Medical libraries have, at least implicitly, the goal of contributing to health. This is the same goal as medical science has. There should thus be no conflict between the way diseases are classified by medical libraries and by medical researchers. There is a danger, of course, that a given social group, say medical researchers or librarians do not serve the users’ goals the best possible way, but tend to serve their own interests. In that case, of course, libraries may choose to include “alternative” medical views. In doing so, however, they are in reality involved in an epistemological struggle about what is proper medical knowledge. In such a struggle information specialists may find allied views within the medical community, within other disciplines and within different kinds of movements, including patients’ organizations and feminist groups. Information scientists do not invent their own criteria from scratch. To a very large extent they relate to points of view already available in the literature. Our working hypothesis should be that different “paradigms” exist in every field of knowledge and that each paradigm serves different goals and interests, for which reason their identification and exploration have the highest priority for Information Science.

Conclusion

This paper has demonstrated that there is a basic correspondence between the views put forward by Feinberg and by Hjørland. Both views are based on epistemological arguments, and both views are connected to the same family of epistemological positions. What seems most urgent is to apply these epistemologies to specific domains. We now have a few adequate domain analytic works in the humanities and social sciences. We need much more investigation, especially the examining of these principles in the sciences.
References


Toward a Theory of Structure in Information Organization Frameworks

Abstract
This paper outlines a formal and systematic approach to explication of the role of structure in information organization. It presents a preliminary set of constructs that are useful for understanding the similarities and differences that obtain across information organization systems. This work seeks to provide necessary groundwork for development of a theory of structure that can serve as a lens through which to observe patterns across systems of information organization.

Introduction
In the wake of the Semantic Web initiative, the rapid evolution of social computing and the growing interest in Second Life, we find ourselves in the midst of what might well be described as the Cambrian age of Information Science. A wealth of new technologies and innovative work practices has generated widespread interest in the development of novel approaches to organizing information, each of which is manifested within a particular context, for a specific group of individuals, in order to address a more or less explicit set of goals and objectives. This focus on the design and implementation of knowledge organization has produced systems that run the gamut from exquisitely crafted, multi-million dollar creations such as the National Cancer Institute’s cancer bioinformatics grid (caBIG) (Saltz et al., 2006) and the National Library of Medicine’s MEDLINE (NLM, 2007) to an emerging smorgasbord of socially-organized information systems such as Del.icio.us (del.icio.us, 2008) and Connotea (Nature, 2008). The upshot is that we are currently confronted by an unprecedented increase in the number and variety of formal and informal systems for knowledge representation and organization.

In light of this explosion of information systems, there is growing need to be able not only to evaluate “new” representational frameworks, such as folksonomies and ontologies, but also to compare these new frameworks with more traditional systems, such as thesauri, subject heading lists, and enumerative classification schemes. While all of these representational systems share the general goal of supporting access to resources, each works in different ways to effect this end. And, although a basic understanding of these systems has evolved within the knowledge organization community, there is increasing need for a theory of structure that can provide a lens through which to compare emerging manifestations; to systematically assess their similarities and differences; to rigorously identify their strengths and weaknesses; and to detect gaps in our own understandings of the utility of these tools for organization and retrieval.

Structure in the Context of Information Organization Frameworks
The structure of a social tagging system, a metadata scheme, or an indexing language must be understood within the framework in which it occurs. The information organization framework itself is comprised of three distinct but interrelated components: the discourse that establishes the goals, priorities and values of the system; the work practices involved in the application and maintenance of the system; and the structure that instantiates both the discourses underlying the framework and the work practices that
make it visible. Thus, it is the work practice(s) and discourse(s) associated with a system that shape apprehension and understanding of its structure. More importantly, it is the discourse(s) and work practice(s) of the domain that will determine the structure of an information organization framework. For example, ontology curation (or engineering) is an information organization framework, and the Gene Ontology (GO) is a specific instance of ontology curation. The discourses revolving around GO reflect the fact that its work practices are focused on representation of the natural (or biological) world; and the structure of GO is therefore informed by this scientific and representationalist focus and the work practices and discourses that follow from that focus.

In order to comprehend the function of structure within an information organization framework — in order to appreciate structure as the product of decisions and priorities established by work practices and discourses within a given domain — it is necessary to begin with a robust theory of structure itself.

Structure is one of those concepts, like information, whose intension is simply assumed. In *Metaphors we live by* (1980), Lakoff and Johnson argue that we comprehend many of the concepts we use by imbuing them with a physical structure that emerges from our day-to-day experiences (p. 59). These “structural metaphors” reflect “systematic correlations within our experience” (p. 61) and allow us to find coherence across diverse experiences in terms of the dimensions of another. In light of their subsequent discussion of experiential gestalts, we may infer that Lakoff and Johnson (1980) refer to the physical components — the “dimensions” — of experience when they speak of structure (1980, p. 83). Nonetheless, Lakoff and Johnson fail to define just what they mean when they speak of structure even though structure-as-metaphor is central to their argument.

Sewell (1992) observes that we frequently conceive of structure as “primary, hard, and immutable, like the girders of a building” (p. 2) — a “thing” that exists independently of experience yet stabilizes and gives shape to it, much like the dimensions to which Lakoff and Johnson (1980) refer. But structure is far more than the sum of the physical components that make up our experience of an entity, an event or a system. In fact, Green (2002) stresses the impossibility of separating the concept of structure from the relationships that link the components of a system. She contends that “[s]tructure and relationships are inextricably interconnected. Wherever structure exists, relationships occur between the components of the structure. Similarly, wherever relationships exist, structure emerges” (p. 73).

Bunge (2003) takes Green’s argument one step further when he contends that not only can the concept of structure not stand alone, but that “there are no structures in themselves” (p. 277). He argues that the apprehension of structure is always dependent on an existing system of relationships and that structure is simply a “property of all systems, whether conceptual or material, natural or social, technical or semiotic” (p. 277). He then defines structure as “the set of all the relations among [a system’s] components, particularly those that hold the system together” (p. 277), and he offers concrete examples to illustrate his argument: “the structure of a sentence is the order of the types of its constituents, such as Subject-Verb-Object in the case of ‘Socrates drank hemlock’; the structure of a theory is the relation of entailment; the structure of a DNA molecule is the sequence of the nucleotides that compose it” (p. 277).

While these intellectual precedents stress the internal nature of structure as a set of components and the relationships between them, it is also important to consider the ex-
ternal aspects of structure and the effects of interaction among structures, work practices and discourses. In his seminal effort to develop a theory of structure for the social sciences, Sewell (1992) defines structure as a duality resulting from interaction between schemas and resources, where schemas are the “fundamental tools of thought, … conventions, recipes, scenarios, principles of action, and habits of speech” (p. 8) and resources are the “manifestations and consequences” (p. 11), the “instantiations or embodiments of schemas [that] inculcate and justify the schemas” (p. 13). Sewell argues that, given that structures demonstrate this duality of origin, “then it must be true that schemas are the effects of resources, just as resources are the effects of schemas” (p. 13): schemas not instantiated in or supported by resources would fade from memory, “just as resources without cultural schemas to direct their use would eventually dissipate and decay” (p. 13). He subsequently defines five axioms of structure that not only follow from the dual nature of structure as both schema and resource but also capture the variability, flexibility and ultimate mutability of structure:

1. **The multiplicity of structures.** Multiplicity ensures flexibility and versatility in that it provides for different discourses and different work practices to adopt and apply different relational models in the design of information systems (pp. 16–17).

2. **The transposability of schemas.** Transposability of schemas allows for different relational models to be applied across a wide array of situations or extended to accommodate the needs of novel work practices or discourses, resulting in different structural forms (pp. 17–18).

3. **The unpredictability of resource accumulation.** The variability arising from the intersection of existing structures, the multiplicity of forms produced by the extension of known schemas, or structures generated as a consequence of resource implementation within competing schemas can not be anticipated or predicted (p. 18).

4. **The polysemy of resources.** Because resources are open to interpretation in different systems, every resource is potentially ambiguous in that it is capable of being re-interpreted within different schemas (pp. 18–19).

5. **The intersection of structures.** The intersection of structures is one by-product of the re-interpretability of resources; but schemas as well as resources can be transposed. As Sewell observes, the “intersection of structures, in fact, takes place in both the schema and the resource dimensions” (p. 19).

In light of these five axioms, Sewell proceeds to define structures as “sets of mutually sustaining schemas and resources that empower and constrain social action and that tend to be reproduced by that social action. … [S]tructures are multiple and intersecting, because schemas are transposable, and because resources are polysemic and accumulate unpredictably” (p. 19). Sewell’s definition can be extended to provide a definition of representational systems as multiple, intersecting and potentially polysemic structures comprised of mutually sustaining discourses and work practices that accumulate unpredictably because they empower and constrain representation and are themselves reproduced, transposed and extended by the act of representation.

**Three Postulates of Structure**

In light of the above discussion, we define structure as a constructed space consisting of a set of internal partitions, each of which is connected to other partitions in the set in a meaningful way, either as a linear sequence (i.e., a continuum or process) or a network of links (i.e., a web) at the lower levels or as a hierarchical or polyhierarchical organization
of part-whole and/or is-a relationships at higher levels. In the context of an information organization framework, a structure is the cohesive whole or “container” created by the establishment of qualified, meaningful relationships among the components, “whether conceptual or material, natural or social, technical or semiotic” (Bunge, 2003, p. 277), which comprise the “bounded space” of the structure.

The definition of structure as it applies to knowledge organization systems requires its own boundaries. These we provide in the form of three postulates:

Postulate 1. The smallest unit of structure is the statement.
Postulate 2. Statements are collected in levels of aggregation.
Postulate 3. The most comprehensive unit of structure is the complex.

A statement is an assertion of a relationship between a resource of interest, an attribute that can be ascribed to that resource, and the value of the attribute as it applies to the subject resource. As such, a statement is a representation of a resource that conforms to the subject-predicate-object format of a simple clause. For example, in the assertion “The title of this paper is Toward a theory of structure in information organization frameworks”, “this paper” is the subject of the representation; “has title” is the predicate that establishes a meaningful relationship between the subject and the object; and “Toward a theory of structure in information organization frameworks” is the value (or object) of the predicate as it applies to the subject of the assertion.

Statements are collected within compound structures that reflect not only the increasingly more sophisticated internal relationships of statements within a structure but also the relationship of the structure itself to the discourse(s) and work practice(s) of the information organization framework with which it is associated. These levels of aggregation consist of records, schemes, systems, and complexes. At the simplest level of aggregation, a record consists of all the statements that have been made about a given resource within a given work practice. For example, the following set of statements, presented in rdf/xml syntax, constitute a record for an electronic resource:

```
<rdf:RDF
   xmlns:dc="http://dublincore.org/2008/01/14/dcterms.rdf#"
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   
   <rdf:Description
      rdf:about="http://www.dlib.org/dlib/june98/scout/06roszkowski.html">
      <dc:creator>Michael Roszkowski</dc:creator>
      <dc:creator>Christopher Lukas</dc:creator>
      <dc:date>1998</dc:date>
   </rdf:Description>
</rdf:RDF>
```

A scheme defines the set of predicates and classes that can be used to make an assertion about a resource. Thus, statements collocated as a record are legitimated by one or more schemes that establish the boundaries of structure by constraining the set of possible statements — the set of all possible relationships that can be established for a resource using that particular scheme. For example, the set of relationships defined as properties in the Dublin Core Metadata Element Set [DCMES] Version 1.1 (DCMI, 2008) along with the Dublin Core Abstract Model (Powell, 2008), prescribes the range of statements that can be made about a resource using DCMES; while it is possible to state the creator, the title and the publisher of a digital resource using properties from the DCMES scheme,
it is beyond the scope of the DCMES scheme to assert, in machine-readable format, the existence of a resource’s mirror site(s).

Classification and categorization schemes used to organize collections of resources, such as Library of Congress Classification [LCC] and Library of Congress Subject Headings [LCSH], generally limit statements about resources to assertions regarding intellectual content. Thus, they prescribe the range of topical statements that can be made about a resource by defining the nature and scope of classes-as-statements. For example, it can be asserted that this paper is about knowledge organization. More importantly, however, when this assertion is situated within an external scheme such as LCC or LCSH, the statement must be interpreted — and additional statements potentially inferred — within the structure of hierarchical or polyhierarchical relationships of classes-as-statements that comprise the scheme.

A system is the instantiation of all records that have been generated within the context of one or more schemes. And while the nature and scope of an individual record is constrained by interaction between the resource and the applicable scheme(s), a system is the result of creating records according to said scheme(s).

It is at the level of the complex that interaction occurs between the non-human resources (the records, schemes, and systems), the discourse(s) of the human agents, and work practice(s). Here the line can be blurred between the act of creating structure and the result of that creative act, the structure itself. Like a small and localized manifestation of Foucault’s épistémé (Foucault, 1980, 1994), the complex is the dermis between the solid state of structure and the gaseous state of discourse.

Mooers’s Method of Descriptors
An example of structure can be found in Mooers’s Descriptor Method. He defines a descriptor as having parts and in so doing defines its structure, the structure of a descriptor. Here we will abstract from Mooers (2003) in order to provide an example of structure.

Our reading of Mooers finds that a descriptor has five parts: (1) idea-element, (2) verbal expression/notation, (3) definition, (4) explicit relationship with a domain, and (5) explicit relationship to descriptor method of information retrieval. There are three intrinsic, non-separable elements (the idea element, verbal expression and definition) and two extrinsic, separable elements (the explicit relationships both to the domain and to information retrieval). These are all elements that can generate their own statements and, when taken in aggregation, form a descriptor record (for “airplanes”, for example, as Mooers illustrates). Here, however, we can only see two explicit statements — the idea element (i.e., the descriptor itself) and the verbal expression of the descriptor; the other parts are not explicit but are derived from context.

This record is not used alone, but is placed within a scheme or system. This scheme/system could be Mooers’s Zatocoding framework (1951) or some other similar framework. Following Mooers’s illustration, we see that “airplanes” is provided along with other descriptors (e.g., “fuselage”, “wheel”, “wing”, etc.), under what Mooers calls a “leading question,” in this case Is a specific component or body studied? (Mooers, 2003). At this point, we can recognize the descriptor as one component of a larger structure that is used in descriptor methodology.

The dual nature of scheme/system is not altogether obvious. The scheme, for our purposes, is the specification of the set of all possible statements (in Mooers’s framework, all possible idea-elements, verbal expressions/notations, etc.), and the system is the actual
instantiation of those statements (in Mooers’s framework, the aeronautical descriptor list from which our example is excerpted, in part). Both scheme and system are necessary for a discussion of structure because many instances of structure in information organization allow for the creation of new statements. We must rely on interaction between the scheme and the system — the intersection of structures — to support this creative work.

Finally this scheme/system is part of what Mooers calls the descriptor method of information retrieval, which is based on a total systems view of the use of information (Mooers, 2003, p. 813). This view posits that a scheme/system will serve a specific and constrained group of uses as well as a specific and constrained collection of documents. This is the aggregation level of the complex where basic structures are deployed in an ever wider structuring of schemas and resources (Sewell, 1992). Key to the descriptor method complex is the well defined purpose of the collection, narrowing the interest in the resources to be retrieved (Mooers, 2003, p. 813).

Beyond complex we reach the outer limit of structure. That limit we call discourse. Discourse outlines the priorities of descriptor methodology in the example from Mooers. Mooers points out the empirical discursive placement of his approach: unlike other representational systems that tend to be implemented consistently and without reference to the context within which the system occurs, “descriptor systems are created at each installation according to a methodology embodying the utmost empiricism” (Mooers, 2003, p. 815). Another aspect of the discourse surrounding Mooers’ work is the focus on the idea-element, rather than terms. As examples in contrast to the Descriptor System, Mooers points to the Uniterm System and the Thesaurus System; it should be possible to speculate the structures of these systems by applying the same framework used to provide an anatomy of the Descriptor System.

**Structure and Its Theory**

Defining structure in this way offers us the descriptive power to compare information organization frameworks from the smallest level of aggregation to levels approaching discourse analysis — from web directories and folksonomies to classification schemes and ontologies. This approach to a theory of structure alerts the analyst to the levels of aggregation at which structure can occur and serves as a touchstone for the interpretation of structure in an increasingly diverse universe of indexing languages. The ability of a theory of structure to identify uniformity across variations in implementation allows us to see where and how domains form design decisions independent of structure. Thus a structural comparison between the Descriptor System and the Uniterm System should provide the domain analyst with a level playing field to negotiate the nuances of the particular domain represented by both systems.

A knowledge organization tool is comprised of many parts, but it has an internal anatomy shaped by the external scaffolding of the work practice(s) and discourse(s) in which it occurs, as Sewell argues (1992). Working toward a theory of structure moves us closer to a systematic understanding of that anatomy and the information organization framework — the skeleton — that is a key component of that anatomy. We assume that we can know the limits of structures in controlled vocabularies (NISO, 2005). Although this assumption is contested by some, if we want to proscribe the novelty of innovative systems and new initiatives (Soergel, 1999) in response to the variability and complexity following on the diversification of practice and tradition, then we must be clear on what grounds. Evaluation of the parts and functions of information organization frameworks stands as one
route to this end. Such evaluation, based on the anatomy of these frameworks, will move us closer to a comparative analysis of utility and a clearer understanding of which structural approaches can provide meaningful alternatives to more traditional frameworks.

References
Knowledge Organization as a Cultural Form
From Knowledge Organization to Knowledge Design

Abstract
By drawing on Lev Manovich’s argument about the database as a cultural form, I argue in this paper that knowledge organization consequently may be considered a cultural form. I use Manovich’s argument to locate an understanding of knowledge organization activity in light of new media. Knowledge organization is not an appendage but rather a prime communication and production form of new media, turning knowledge organization into knowledge design. This implies moreover that knowledge organization is not only an activity ensuring storage and retrieval but is also media design as well. The theoretical consequence of this is to engage in ‘info-aesthetics’.

Introduction
In his book *The Language of New Media* (Manovich, 2000), Lev Manovich argues that the database is the new cultural (or symbolic) form in our culture. Before that, Manovich claims, the narrative was the dominating cultural form. The database makes it possible to (re)present and (re)configure culture in ways not possible before the computer. It means that our way of approaching and using knowledge and information is changed (once again) due to a particular medium. In a previous article (Andersen, 2002), it was argued how various forms of media (orality, writing, printing and the computer) influence on our concept of knowledge organization in library and information studies (LIS). This argument was two-fold: 1) Knowledge organization lacks an awareness of media and how they historically shape the way societies and cultures store and organize knowledge, and 2) Media not only shape how knowledge is stored and organized, media also organize knowledge themselves and that this aspect is to a large extent ignored in knowledge organization research.

In this article, I want to further develop the argument presented in Andersen (2002) by taking a specific point of departure in Manovich’s argument about the database as a new cultural form. I want to examine what such an understanding of digital media brings to knowledge organization research. I believe that we still do not pay enough scholarly attention to digital media in knowledge organization research and it is this I want to address here through Manovich’s argument. It is more imperative than ever to pay attention to digital media as the practices of storing and using knowledge and information are not separated with digital media. Moreover, digital media abound and produce what we in the knowledge organization research community consider as knowledge organization activities.

Manovich’s discussion is a challenge to and informing for knowledge organization research exactly because it takes a view on a specific practice (the database and its collection(s)) that bear a resemblance to what is studied in knowledge organization research and elevates this to cultural and aesthetic level.

The structure of the article is as following. Having made a suggestion as to reconfigure knowledge organization, I move on to outlining Manovich’s argument about the database as a cultural form. From here I discuss the implications for knowledge organization research.
Reconfiguring knowledge organization

I suggest we broaden our understanding of knowledge organization so as not to exclude social practices and human activities that do organize knowledge but in a different form and on a different level than indexing, cataloguing and classification (cf. Andersen & Skouvig, 2006). This will bring knowledge organization theory in much closer contact with sociological and cultural understandings which is needed if knowledge organization research intends to have a social and cultural relevance. Therefore, I suggest we understand knowledge organization research to be concerned with knowledge organizing activities as they are practiced and mediated by societies, humans, information and communication technologies, texts, media and genres and the relationship between these. Thus, knowledge organization research has (at least) two levels: 1) The study of problems related to the organization and representation of texts in various forms of information systems, and 2) The study of social and cultural practices and discursive activities understood as knowledge organizing categories. This opens up for including the study of, for instance, how texts, genres and media organize knowledge, a perspective not that much present in knowledge organization research.

The database as a cultural form

The notion of the database as the new cultural form could lead one to conclude that knowledge organization is a new cultural form and to some extent it may be. The general public is more than ever in touch with and uses more than ever forms of knowledge organization such as search engines, folksonomies, or wikipedias. The content of these systems is to a very large extent produced by the general public and, consequently, these forms of knowledge organization are not restricted to particular professional domains.

But when stating that the database is the cultural form of our time, how does Manovich arrive at this statement? To claim that a particular instance of computer technology (the database) represents a new cultural form may sound not only deterministic but may also be to ascribe a cultural significance to the database more than it actually has or is capable of having.

Manovich’s argument is sophisticated drawing on both the database as ‘just’ a piece of technology and on what he considers as ‘old’ cultural forms: the cinema (and the novel):

After the novel, and subsequently cinema privileged narrative as the key form of cultural expression of the modern age, the computer age introduces its correlate — database. Many new media objects do not tell stories; they don’t have beginning or end; in fact, they don’t have any development, thematically, formally or otherwise which would organize their elements into a sequence. Instead, they are collections of individual items, where every item has the same significance as any other. (Manovich, 2000, p. 218; my italics)

Such is Manovich’s start out. Of course, representing a list of items is not a new activity in human culture. Jack Goody, for instance, makes the argument that the list is a technology that is not simply a reproduction of oral discourse because the list as a formal and abstract device does not represent speech directly (Goody, 1977, p. 80). But the computerization of society and the ontology of the computer invite the database form (or the list form) and introduce it as an everyday form.

In continuation of this, Manovich (2000, p. 217) claims that “Information access has become a key activity of the computer age; it is a new category of culture. From the user’s point of view, new media appear objects as databases because users can perform various actions with them like viewing, navigating, and searching:
The user’s experience of such computerized collections is, therefore, quite distinct from reading a narrative or watching a film or navigating an architectural site. Similarly, literary or cinematic narrative, an architectural plan and database each present a different model of what a world is like. It is this sense of database as a cultural form of its own which I want to address here. (Manovich, 2000, p. 219)

Almost echoing the famous McLuhan-statement ‘the medium is message’ (McLuhan, 1964), Manovich tries to show how the database has evolved into a cultural form. That is to say, how culture has moved into the database and then turning the database into more than just a piece of computer technology. About this Manovich (2000, p. 225) says that what makes the database a cultural form is its way of representing the world as a list of items and not as a narrative. Thus, what makes the database a cultural form is what on the surface seem to be the ‘innocent’ characteristic of the database. This claim about the database is of course not to say that the database as a physical and technical object is a cultural form but rather what the database is capable of doing (i.e. representing a list of items), or what users are capable of doing with a database, is what constitutes a cultural form. With this, Manovich claims that representing the world as a list of items has become a dominant form of culture in the digital age.

Thus, Manovich brings us an understanding of the database as more than a technological phenomenon but as an articulation and materialization of human culture. This insight is what I want to relocate in knowledge organization research as it provides ground for an understanding of knowledge organization as a specific cultural practice; a specific way of handling the products of human culture and human activity. In short, what I am trying to argue in the following is that, paraphrasing McLuhan, ‘knowledge organization is the message’, meaning that the activity of organizing knowledge is not just a ‘technical operation’ performed on e.g. documents or knowledge, but that this activity is an everyday cultural practice and a mode of production in our culture.

Knowledge organization as cultural form: From knowledge organization to knowledge design

Human societies have throughout history created diverse forms of communication and forms of culture and knowledge organization activities can be considered a response to such forms. Given that the database is a form of knowledge organization implies that knowledge organization is part of a new cultural form, a new common cultural practice. We can observe many kinds of such new cultural forms.

A website does in most cases contain a collection of links to either internal pages or external pages: “There is hardly a Web site that does not feature at least a dozen links to other sites; therefore, every site is a type of database”, making indexing a new hobby (Manovich, 2000, p. 225). Search engines practice and remediate to some extent the old function of enumerative bibliography in digital culture. Digital libraries and archives organize collections of texts and hence act like metatexts themselves (Dalgaard, 2001). Folksonomies and the practice of tagging are expressions of cultural forms of communication and sharing of knowledge. Weblogs contain archives of previous postings and in some cases search fields. Digital scholarly journals perform archival activities at the same time as they appear as journals. Thus, whereas many these activities in print culture were separated, they are now integrated in digital media. Moreover, what matters here is that these cultural forms on the surface are not considered databases and rightly so. But the way they perform is to a large extent shaped by the database form. The very activity of these cultural forms in digital media could not have been accomplished without the un-
derlying database form. That is, our actions with these cultural forms are different from cultural forms relying on for instance the narrative or linearity in their structure. In digital media, narrativity or linearity is one option among many.

The very practice of collecting links, searching, tagging and archiving in digital media performed by all sorts of people suggests that these are everyday activities exactly because digital media invite these practices due to the database ontology of digital media. Thus, not only represents the ubiquitous database form a new cultural form, it also seem to remediate older forms of knowledge, namely the list and the encyclopedia. But there is a difference. It is not a matter of seeing databases as storage and retrieval forms only but primarily to recognize databases as particular forms of design and communication in digital media. In this way, knowledge organization becomes knowledge design. It is the design of knowledge for action.

Thus, we can understand knowledge organization practices in new media as cultural forms, as modes of designing and communicating culture. Not as something technically inherent in digital media but something digital media invite us to do in much the same way writing invites writing down myths, producing novels or communicative stability. The point is that we could have chosen to see these practices as anything else than cultural forms. But by viewing them as cultural forms we subscribe to them shared symbolic meaning and shared modes of action that make symbolic forms act or perform in specific ways in human activity.

**Knowledge organization as cultural form: The implications and perspectives for knowledge organization theory**

The notion of knowledge organization as a cultural form implies that we must start to understand knowledge organization as mode of designing and communicating culture and knowledge. Is that not what knowledge organization always has been concerned with? Yes and no. Knowledge organization has been very much concerned with communicating culture and knowledge by means of, for instance, classification systems. But knowledge organization has not been viewed as a production and aesthetic form. In new media the database form is a dominant creative production form:

> The database becomes the center of the creative process in the computer age. Historically, the artist made a unique work within a particular medium. Therefore the interface and the work were the same; in other words, the level of an interface did not exist. With new media, the content of the work and the interface become separate. It is therefore possible to create different interfaces to the same material. (Manovich, 2000, p. 227)

Understanding knowledge organization as a cultural form, as knowledge design, invites a rethinking and a reconfiguration of knowledge organization theory and practice. Knowledge organization theory should be reconfigured along the lines of what Manovich calls “info-aesthetics”:

> ...a theoretical analysis of the aesthetics of information access as well as the creation of new media that “aestheticize” information processing. (Manovich, 2000, p. 217)

Given that information access is a ‘key category of culture’ (Manovich, 2000, p. 217), it is necessary to recognize the ‘culturalization’ and ‘aestheticization’ of knowledge organization that new media bring with it. This means that our present vocabulary and approaches to knowledge organization are limited. We must engage in discussions about how and with what means knowledge is designed in new media, what actions are to be accomplished with knowledge design. To some extent parts of this is not new to knowledge organization theory as human-computer interaction and information visualization
have also been contributing. But with Manovich we get a vocabulary and an understanding that situates interfaces and knowledge design in a larger media-historical perspective. The approach is humanistic in the sense that media and media development are seen as cultural products and cultural developments together with media-technological developments. In knowledge organization theory we are not used to see knowledge organization theory in such a perspective partly because we may have had a too narrow understanding of what knowledge organization is or could be.

Understanding knowledge organization as a cultural form broadens the domain of knowledge organization both in its theory but also in its practice. On a theoretical level, we gain insights from media theory as to how human experience and culture has been and is shaped and structured by various forms of media, with the computer seen as the latest enlargement of our media-cultural matrix (Finnemann, 2001; McLuhan, 1964). Various new media do not replace older media but develop alongside existing media and stabilize and adjust themselves in the media-cultural matrix. The content of any new media is always older forms of media (McLuhan, 1964). The computer is just a medium whose content is defined by older media such as radio, television, mail, books. But due to its digital status the computer is capable of treating older media as variables in the computer medium (Finnemann, 2001). Thus, we can say that the computer treats older forms of media as items to be structured. This leads to the practical outcome. Knowledge organization in new media is not only about ordering a list of items for retrieval. It is about seeing knowledge organization as a creative and aesthetic practice. Knowledge organization in new media is to produce a work of art in the sense that making access to the work is a design process (knowledge design) and not a process of following or applying rules, techniques, or standards. This suggests awareness towards how to materialize and communicate digital culture. Such materialization and communication of digital culture is performed by many people as part of their professional and everyday life. Thus, this strikes back on knowledge organization theory and research. We have to take seriously the activities in which knowledge design is performed by people and gain knowledge about what kinds of shared symbolic meaning people ascribe to their knowledge design activities. In a discussion of what constitutes rhetorical genre, Carolyn Miller (1984) argued that such an understanding of genre should not be located in the form or substance but on the actions the genre is used to accomplish in people’s everyday communicative practices (Miller, 1984). Likewise, I suggest for knowledge organization theory that we should not be looking ‘closed’ understandings of knowledge organization and the form and substance of knowledge organization but on the actions knowledge organization perform for and by people in everyday digital communication practices. This will provide us with a window as to how and to what extent knowledge organization as knowledge design is a cultural form.

Concluding remarks
Outlining an understanding of knowledge organization as a cultural form supplies us with a means for situating knowledge organization activities in light of new media. The challenge is to start to see knowledge organization as a form of media itself in digital culture. That is, knowledge organization is not an attachment to other activities communicating and designing knowledge. It is a knowledge design activity itself. In digital media the database is the prime communication and production form. Building on Manovich’s notion of ‘info-aesthetics’, knowledge organization research opens for an orientation that stresses to understand culture and identity in knowledge organization research as a matter
of aestheticizing knowledge design activity and to understand this activity as media creation as well. Aesthetics, then, brings a new dimension to knowledge organization theory (knowledge design theory); a dimension that provides knowledge organization theory with a view that sees knowledge design as creative digital art.

References
Origins of the Main Classes in the First Chinese Bibliographic Classification

Abstract
The aim of the paper is to provide an improved understanding of the classification applied in the Seven Epitomes (Qilüe), the first documented classified library catalogue in China (completed in the first century BCE). Chinese bibliographers have suggested that Liu Xin, the compiler of the catalogue, followed the Principle of Classicist Values, state of scholarship, literary warrant, and ideas of yin/yang and the Five Phases to devise the six-fold classification. By applying a multidimensional framework constructed for a large-scale research project, the author re-examined the origins of the six main classes in the catalogue within its own social, cultural, and political contexts. Issues highlighted for discussion include the concept of “discipline”, the limitation of the classification in relation to literary warrant, and the motives of intellectual control and social engineering.

Introduction
This paper will examine one of the fundamental aspects of traditional Chinese bibliographic classification: the origins of its main classes. Chinese bibliographic classification maintained its own tradition for about two thousand years, approximately between the first century BCE and the early twentieth century CE before the influence of Western bibliography came to knock on China’s door. This knowledge organization tradition has attracted limited attention from classification theorists in the West. Although classification is an important branch in library and information science, its published literature in English has focused mostly on theory and practices emerging from the Western traditions. In basic English texts introducing classification, frequently mentioned thinkers who have significantly influenced classification theory and schemes include Aristotle, Linnaeus, Francis Bacon, and Melvil Dewey (e.g., Broughton, 2004). The only Asian classification theorist known to English-speaking researchers and practitioners is Shiyali Ramamrita Ranganathan, who received his library science education in London, England, in the early twentieth century. Little is known to date about bibliographic classification efforts outside Western cultures. Awareness of this deficiency in the literature about classification has recently been heightened as more scholars write about the Western, Christian, and male bias in commonly known classification schemes like the Dewey Decimal Classification (Olson, 2002).

Published works in English about traditional Chinese bibliographic classification are few; the most substantive are two journal articles published more than fifty years apart. The earlier of the two articles describes the historical developments of Chinese bibliographic classification through the first half of the twentieth century (Tsien, 1952). The second work, in addition to the history, provides a general discussion of the cultural values reflected in traditional Chinese bibliographic classification, mostly summarizing the views expressed by Chinese bibliographers (Jiang, 2007). This kind of sporadic and sketchy treatment of a two thousand-year classification tradition in the literature leaves plenty to be desired. Even some of the most fundamental aspects of Chinese bibliographic classification remain enigmatic to scholars in the West. In view of this deficiency, I have started a large-scale project to research traditional bibliographic classifications in imperial China.
Figure 1. Comparison of the main classes in two schemes²

For a more in-depth analysis, I will highlight one particular case in the study reported here—the main classes of the classification applied in the *Seven Epitomes* (Qi lüe 八略), the first classified library catalogue, completed between 6 and 1 BCE, in China. This scheme had six main classes (see Figure 1).² Since then a number of variations have developed: four-fold, five-fold, seven-fold, nine-fold, twelve-fold, etc. All of these, nevertheless, followed the same basic structure of the six-fold scheme in the *Seven Epitomes* with individual classes being merged or divided. Figure 1 gives a comparison of the classes in the *Seven Epitomes* and those in another scheme devised much later (the end of the eighteenth century), the *Si ku quan shu* “zong mu” 四庫全書總目 (the catalogue of a collection of books titled *Si ku quan shu*). Studying the origins of the main classes in a scheme that started a tradition will shed significant light on the tradition.

The study will begin with a brief introduction to the *Seven Epitomes*, followed by a review of the literature about the origins of the main classes in the *Seven Epitomes*. By applying a multi-dimensional framework, the next section expands the discussion by highlighting a number of issues not fully considered before.

The *Seven Epitomes* and Its Classification

The *Seven Epitomes* is said to be compiled by a distinguished scholar Liu Xin 劉歆 (53 BCE–23 CE). As a by-product of a collation project commissioned by the Emperor, it was the catalogue of all collated books housed in the libraries of the Inner Court at the time. Although it is no longer extant, Chinese bibliographers believed that its majority entries (in a much abridged form) and its original classificatory structure have been preserved in the “Bibliographic Treatise” of the *History of Former Han Dynasty* (*Han shu* “yi wen zhi” 漢書藝文志, compiled about a hundred years later and referred to as the “Treatise” hereafter). Scholars estimate that there were more than six hundred annotated entries in the *Seven Epitomes* arranged according to a carefully designed classification.

The title of the catalogue seems to suggest that its classification consisted of seven epitomes (classes). However, the “Treatise” included only six classes (without “Ji lüe” or the Collective Epitome). Since the *Seven Epitomes* is no longer extant, scholars have not been able to come to a consensus about the nature and content of Ji lüe. One speculation that has been widely accepted is that Ji lüe was the collection of brief summaries now

---

1. The Chinese terms for the classes in the first scheme are “Liu yi lüe” 六藝略, “Zhu zi lüe 諸子略”, “Shi fu lüe 詩賦略”, “Bing shu lüe 兵書略”, “Shu shu lüe 術數略”, and “Fang ji lüe 方技略”; and the classes in the second scheme are “Jing 經”, “Shi 史”, “Zi 子”, and “Ji 集”. The English translation of the six classes in the former is from Lewis (1999), and the English translation of the four classes in the latter is from Tarsala (2001).
2. In this paper, pinyin is used for romanizing Chinese names and titles.
seen at the end of each of the six main classes and their divisions. Nevertheless, no one disputes that the classification in the *Seven Epitomes* was a six-fold scheme.

There are six classes and divisions in the *Seven Epitomes*:

1. **Liu yi lüe** (Epitome of the Six Arts) consisted of nine divisions, including one for each of the Six Classics (*Odes, Documents, Rites, Music, Changes*, and *Spring and Autumn Annals*), *Analects* of Confucius, *Book of Filial Piety*, and philology.
2. **Zhu zi lüe** (Epitome of the Masters) consisted of ten divisions, including nine major affiliations of thought commonly known during the Warring States and an added affiliation of Novelists.
3. **Shi fu lüe** (Epitome of Lyrics and Rhapsodies) consisted of five divisions, including three styles of poetry and two other genres.
4. **Bing shu lüe** (Epitome of Military Texts) consisted of four divisions (tactics, terrain, yin/yang, and military skills).
5. **Shu shu lüe** (Epitome of Numbers and Divination) consisted of six divisions, including astronomy, chronology, five phases correlative elements, divination, miscellaneous fortune-telling, and geomancy.
6. **Fang ji lüe** (Epitome of Formulae and Techniques) consisted of four divisions, including medical classics, pharmacology, sexology, and longevity.

**Related Literature**

Researchers of the *Seven Epitomes* face two obstacles. First, the catalogue was lost long ago. Some scholars estimate that the catalogue disappeared between the late ninth century and the first half of the tenth century, basing their assertion on the fact that the *Seven Epitomes* was not listed in any extant bibliographies compiled in Northern Song Dynasty (960–1127) and after. Being part of the *History of Former Han Dynasty*, the “Treatise” has been able to survive. Today researchers of the *Seven Epitomes* must use the “Treatise” as the basis for analysis, supplementing it with fragments quoted in other sources. Studies focusing on the “Treatise” alone without referencing the *Seven Epitomes* are also useful for understanding the latter.

The second difficulty in studying the classification of the *Seven Epitomes* is the extended time lag between the catalogue’s creation and any discussions of its fundamental design principles. To date there are no known writings by the compiler, Liu Xin, articulating the principles or ideas behind it. Besides, bibliography was not a significant field of study until Qing Dynasty (1644–1912), so substantive discussions of bibliographic or classificatory principles only emerged hundreds of years after the compilation of the *Seven Epitomes*. Researchers thus need to be mindful of this time lag when reading the literature on the *Seven Epitomes*.

According to a long-held belief, the dominating factor in determining the six main classes and their sequence in the *Seven Epitomes* is Ru classicism (also known as Confucianism). We may call it the Principle of the Classicist Values. It is said that Liu Xin, like other classicist scholars, considered the Six Classics to be the canons for modeling government and personal conduct and thus devised the foremost class the Epitome

3. The Six Classics, sometimes referred to as the Six Arts, are later known as the Five Classics, for the *Music* was lost long ago.
4. Several nineteenth-century scholars reconstructed the *Seven Epitomes* based on the “Treatise” and fragments from other sources. Among those versions, the one compiled by Yao Zhenzong (1978) is most authoritative.
of the Six Arts to include these six texts and their interpretations and commentaries. By the same token, “Numbers and Divination” and “Formulae and Techniques” (together encompassing several fields of science and technology) were at the end of the classificatory structure because classicists viewed them as the least significant of all branches of knowledge (Tan, 2003).

As suggested by Chinese scholars, the second factor that influenced the divisions of the main classes in the Seven Epitomes was the state of scholarship (Zuo, 2004). It was stated in the imperial decree ordering the undertaking of the collation project in six broad categories—exactly the same as the six classes in the Seven Epitomes. For some, this statement proves that the six-fold categorization was not a creation of Liu Xin; rather, it reflected the common divisions of scholarship in Former Han. It also means that the scheme in the Seven Epitomes was both a bibliographic classification and a classification of knowledge.

In comparing the main classes in the Seven Epitomes and its successors, the issue of literary warrant, somewhat related to the state of scholarship, comes up frequently. A popular view asserts that Liu Xin considered literary warrant in balancing the number of texts in each of the six classes. In particular, the Seven Epitomes had no “History” class (see Figure 1) because there were comparatively few texts of history available at the time. Military texts were numerous during Han times due to frequent military conflicts that lasted for hundreds of years before then; thus a separate class for these works was justified. This view has been challenged (Wang, 1998). See the “Important Issues for Discussion” section for more discussion.

Another camp maintains that Liu Xin was under the influence of the theories of yin 陰 and yang 陽 and the Five Phases (wu xing 五行), popular in Han times, in creating the six classes (Zhang, 1994). In this belief, it is said that the classification of the Seven Epitomes was an attempt to return texts to a unified whole with the Six Classics as the origins of all other written texts (Lewis, 1999). Some of the scholars holding this view are convinced that Liu Xin forged the text of the Zhou li 周禮, the second part of the Book of Rites (one of the Six Classics), taking the same philosophical approach in which the ideal state administration consisted of six offices, with one of them being the Chief Minister (tai zai 太宰) to lead the rest.5 The fact that the six categories were specifically mentioned in the imperial decree cited above makes it unlikely that Liu Xin invented the six classes.

Methodology
In my large-scale research project on bibliographic classifications in imperial China, I have established an analytical framework as an attempt to bring new insight into this area of study, which will be continuously refined throughout the project. This framework is composed of (1) the text of a classification (the reconstructed version in the case of the Seven Epitomes); (2) the biographical information about the classificationist; (3) the history of the government, especially the makeup of its bureaucracy and ideology (because many targeted bibliographies were commissioned by the throne); (4) the intellectual history; and (5) the technological conditions at the time. Such analytical and interpretive framework offers a multi-dimensional approach to better contextualizing data that is anticipated to have two benefits in this historical research. First, it will be useful for identi-

5. The dispute about whether the Zhou li 周禮 was a forgery created by Liu Xin has lasted for more than a century. Today judging from recent archeological discoveries, scholars generally do not support the accusation.
fying various intentions and considerations, explicit as well as hidden, in the design of the classification. Second, situating data in the original cultural context helps to avoid interpretive errors made by imposing an epistemology of a different culture in analysis.

Another method applied to dealing with historical texts in this study is known as triangulation of sources—identifying more than one account or explanation of the same event in multiple and preferably independent sources. Due to the age and current state of the Seven Epitomes, this method is especially necessary for achieving a rigorous analysis.

**Important Issues for Discussion**

As this research is ongoing, I will bring up for discussion the following issues that are of crucial importance in classification theory. The first issue to address is the concept of “discipline”. In Han China when the Seven Epitomes was compiled, the Chinese perceived and studied the world in ways very different from those structured according to disciplines as known today. Tarsala (2001) points out this problem in the previous English translation of the four classes in the Si ku quan shu “zong mu” (Figure 1). To avoid this type of misunderstanding, she proposes to use “Histories” and “Masters” to replace “History” and “Philosophy”, the latter two commonly referring to two disciplines in the West. In the above “Related Literature” section, I mentioned Wang’s disagreement on the role of literary warrant in the omission of a “History” class in the Seven Epitomes (Wang, 1998). He enumerates many texts scattered in various classes in the Seven Epitomes that were writings of history, enough to justify a separate class to accommodate them. The reason for not having a history class, he argues, is that history as a field of study was in its infancy in Former Han, and Liu Xin followed the scholarly landscape understood in pre-Han times to devise his classification. Quite interestingly, both Tarsala and Wang challenge the concept of discipline; the former suggests to rid of it from the study of Chinese bibliography and the latter defines it as what the ancient Chinese conceived of. This issue is at the heart of bibliographic classification and warrants additional scrutiny.

While literary warrant has been discussed, bibliographers mostly associate it with the balance among the classes. We must make it clear that the Seven Epitomes was not a comprehensive record of all writings available at the time; its coverage limited the scope of its classification. In fact, the books included in the Seven Epitomes were only those in the imperial libraries of the Inner Court that were part of the collation project. Collections in other government agencies were not part of the collation project at all and thus were left out of the Seven Epitomes. Writings on laws and elementary mathematics, for example, were excluded because they were the responsibility of other government agencies (Gu, 1989; Liu, 1982). Therefore, the claim that Liu Xin intended to treat all texts in the Seven Epitomes as a unity seems problematic and calls for further investigation.6

To understand the classification of the Seven Epitomes, it is also necessary to elaborate on the motives of intellectual control and social engineering. Han was the first dynasty that was able to rule China with a powerful and long-lasting central government. Its emperors believed in intellectual control. But the earlier emperors in the first sixty years or so were supporters of Daoism (as well as Legalism to some extent). Emperor Wu 武 (r. 141–87 BCE) started a new policy to fiercely promote classicism as the state-sanctioned

---

6. Writings about laws were very important government documents. There are historical records showing that law texts were among several major categories in another earlier collation project. Thus, Liu Xin, being a central government official and renowned scholar, could not be ignorant of their existence.
ideology. Approximately one hundred years later, then Emperor Cheng 成 (r. 32–7 BCE) commissioned a collation project as part of its effort to establish intellectual authority dominated by classicism (Lewis, 1999; Nylan, 2001). In the process of collation, the participating scholars, Liu Xin included, and subject experts sorted through texts written on bamboo and wooden slats, many having fallen out of order, deciphered fading characters, and compared various versions. They then created a finalized version of each text, some of them new anthologies that have become the books as we currently know. The adoption of the classicist values by Liu Xin in reformulating texts and creating the catalogue and its classification was not simply to reflect the cultural values as they naturally occurred or developed. Clearly the ruling class played a strong hand in shaping the values.

Conclusion
Several issues emerged in re-examining the origins of the main classes in the Seven Epitomes by applying the aforementioned multi-dimensional framework for analysis. First, the classificationist’s conceptualization of “a field of study” in the original context is central to a better understanding of the classification itself. This is especially critical if the bibliographic classification in question is closely associated with classification of knowledge. Second, literary warrant translated into real limitations for the Seven Epitomes in terms of its coverage of written texts. It is advisable for researchers to keep in mind such limitations in their study of the overall classificatory structure of the catalogue. Third, the political overtones in the Seven Epitomes stemmed from the fact that the catalogue was a by-product of a government-sponsored collation project. Placing it in its original context, the analysis helped to illuminate the catalogue’s political and social functions that are as important as those of a utilitarian nature.

The above discussions gave early indication of the effectiveness of the proposed analytical framework. In the literature on Chinese bibliographic classifications, the lack of such framework has lead to erroneous interpretations, and even unwarranted opinions stating that Chinese bibliographic classification was underdeveloped. The observations made in this paper are not meant to be final. In the ongoing research project, we will continue to study the theoretical basis of the Seven Epitomes and a number of its successors within the framework.

Acknowledgements
This project was made possible by a Fulbright Scholarship Research Award. I wish to thank Peking University, Department of Information Management for hosting me during my Fulbright study in 2006. Additionally, I acknowledge the significant contributions made to the overall research project by Dr. Lan Wen-Chin of the National Taiwan University.

References


The Creation of Keysigns
American Sign Language Metadata

Abstract
This paper reports preliminary results from a pilot test of the creation of a folksonomic gestural taxonomy for sign language indexing and retrieval. Skilled sign language interpreters and deaf participants were asked to create sign language metadata or ‘Keysigns’ that they would assign to classify topics presented by three deaf scientists during a day-log workshop. Although their Keysigns demonstrate a high degree of content conformity, the physical signing itself lacked consistency. Comments made by participants revealed that signed metadata was not a commonly understood concept and that the exercise was cognitively challenging. The paper concludes with suggestions for ways to make the creation of folksonomic Keysign metadata easier from cognitive and physical perspectives.

Introduction
American Sign Language (ASL) is the most popular sign language in Canada and the United States and Langue des Sourds du Québec (LSQ) is the second-most popular sign language used in Canada (formerly la langue des signes Québécoises). In many cases, these languages are the first language and main means of communication for deaf individuals. Sign languages such as ASL are complete visual-spatial languages, governed by their own rules for syntax, grammar and punctuation (Stokoe, 2001). As such, they are not direct translations of any specific spoken languages. The signs express concepts and not necessarily individual words per se, similar to how groups of words in a written language are used to express concepts.

Signs used in sign language can be defined by the following general characteristics: handshape (including finger orientation, and wrist position); palm orientation (palm position in 3-dimensional space or pitch, yaw and tilt); location of hands in space and relative to each other and the body; direction and speed of movement of the hands, arms, wrists and fingers in 3-dimensional space; and non-manual signals such as body position and facial expression (Tennant & Brown, 1998). While these characteristics may be helpful in categorizing signs, and constructing static or animated versions of signs, they are extremely difficult to disambiguate in recordings of sign language using automatic recognition or even manual systems for recognition. In order to build an indexing system for sign language, a simpler, more pragmatic approach that users can control and that is based in sign language itself is required.

Spoken or written language is often a second language for deaf individuals and consequently, literacy in English or French for this group is limited. In a study of the literacy levels of deaf and hard of hearing high school students, Holt (1997) found that deaf ASL-speaking teens graduating from high school have an average English literacy level of grade four. English text and translated dialogue (from Web, television or digital video content) are much less accessible than if the language of presentation was ASL. Deaf persons therefore may have considerable difficulty accessing and taking advantage of the Web, an important possible source of information, services and entertainment for them.
These issues can be compounded for a person whose first language is ASL or LSQ because typical training materials, and on-line technologies are based on a written language such as English that is much less accessible. It is important to provide on-line content in ASL and other sign languages as this is the optimal method of communication for many deaf persons.

However, expressing ASL on the web requires a movement-based display technology such as video or animation because of the visual-spatial nature of sign language. As a result of this requirement, written text is often used for explanations of the sign language video appearing on a website, for navigational elements such as button labels and menu items, and to contribute unique content that is different from that expressed in the sign language video. In addition, this sign language content is not a searchable entity — users cannot search content expressed within the sign language video. Access to most collections of sign language video has thus used textual surrogates for items in the collection as well as textual queries.

One solution proposed by Richards et al., (2004) is SignLink Studio, a web-based editor that has been developed to allow web authors to create web pages using sign language video. Hyperlinks can be added to video material so that users can navigate between multiple videos without the need for text (although text is optional). Screen shots and prototype examples can be seen at www.aslpah.ca (also featured on Canadian Heritage’s culture.ca website in May 2006).

Hyperlinking alone will not make signed resources on the web accessible, and the information will only be useful if they can be searched in some way. As the corpora of videos and signed websites increases, a method of organizing using sign language will be necessary. Access to these materials is currently challenged by the lack of sign language metadata that can provide a summary of the resources, and facilitate navigation and retrieval in the native language of the resources and of the users.

Current approaches to sign language indexing have focused primarily on textual access mechanisms, most notably the ISLE Meta Data Initiative (IMDI) sign language metadata element set created in cooperation with the ECHO project (Crasborn & Hanke, 2003; ECHO, 2007). At the same time, recent advances in video processing and image recognition have also spurred the development of gesture recognition systems for sign language recognition and translation (Bowden et al., 2004; Bowden & Sarhadi, 2002). Missing from these approaches is the creation of visual metadata or “Keysigns” that may be used as surrogates for browsing and retrieval in the same language as the resources themselves. Surrogates function not only as attributes against which a query may be matched, but also provide support for browsing, navigation, relevance judgments and query reformulation. In order to be effective, surrogates must convey the content of the original in such a way that users will make the same distinctions between surrogates that they would make between full documents.

However, image centric metadata is not readily or easily implemented due to the complexity of analyzing and interpreting image centric data. Although there is considerable research effort in the video/visual analysis area, there are still no optimal solutions particularly for visual/spatial language recognition. Additional complexity is introduced as there are large variations in the production of signs by individual users. For example, people can generate “sloppy” sign or slur them together analogous to a “sloppy” or slurred pronunciation of a word, or users can use more or less space in which to generate a sign, or can sign faster or slower. The ideal computer-based Keysign system that would au-
tomatically recognize and disambiguate Keysigns in sign language recordings requires sufficient intelligence to accommodate for these individual differences in sign language production.

This paper is concerned with the problem of indexing digitized sign language videos for information retrieval. Of primary concern is the provision of sign language surrogates to provide access to sign language resources. We want to develop a set of non-text based keywords (Keysigns) created using sign language that can be used to access on-line sign language content. The research and development of Keysigns is based on an image-centric approach rather than a text-based approach to metadata. The video-based Keysigns will be a set of user-selectable Keysigns that have been developed by users; creating a folksonomic gestural taxonomy. Users can then use/select one or more Keysigns that appropriately describe their video content from this set. User-determined categorization strategies are becoming increasingly commonplace in text-based retrieval systems, and folksonomic tagging has the ability to support the unique language and information needs of specific user communities (Smith, 2007).

The following sections of this paper report preliminary results from a pilot test of the creation of a folksonomic gestural taxonomy for sign language indexing and retrieval. When it is completed, the taxonomy will be a collection of signed concepts to support browsing, navigation and retrieval.

**Methodology**

A full day workshop was held that focused on challenges faced by deaf academics, academic sign language interpreters and the use of technology in academic interpreting. The workshop consisted of eight signed presentations on various topics including examples of research conducted by deaf academics and the role of academic interpreters. The presentations were videotaped and later digitized. Three presentations were selected as content for this study. The first presentation, on the topic of working with interpreters in higher education, was given by a hearing presenter speaking in ASL. The second presentation, given by a late-deafened person speaking in CASE, was on the topic of working as a deaf academic in higher education. The third presentation, on the topic of genetics, was given by a pre-lingual deaf presenter speaking in ASL/CASE.

Participants in the study were recruited from three groups: interpreters who attended the full day workshop, interpreters who viewed the videotaped presentations, and deaf students who viewed the videotaped presentations. Participants in each group were asked to complete a brief demographic survey and were then asked to provide up to five Keysigns to represent the content of the each of the three presentations. Participant generated Keysigns were videotaped for later analysis.

Of the nine participants in the study, four were deaf, and five were non-deaf professional sign language interpreters. The sign language interpreter community is highly feminized, and all interpreters participating in the study were female. Of the nine participants overall, only three participants were male. Participant ages ranged from 30–59. All but one participant had a college degree.

A number of challenges emerged in the eliciting of Keysigns from participants. First, there were a number of concepts which lacked an ASL equivalent including the concepts of Metadata and Keysign. We chose to represent the concept of ‘Keysign’ using a sign whose closest meaning in English was “main point sign.” In ASL conversations when we signed to deaf people about Keysigns, we fingerspelled the word first, then explained...
what it meant, and its purpose using signs indicating “important/crucial point”. Nevertheless, deaf participants expressed frustration with the activity of creating signed metadata because the concept was so unfamiliar. In contrast, when explaining these concepts in English to non-deaf participants (interpreters) it was relatively simple to explain the concept of Keysigns and to make comparisons with the concept of Keywords. Indexing and categorizing written language has been a common activity for many years and it has thus become a commonly understood concept. Yet, for sign language users, this was a new way of conceptualizing the language and new concepts for describing it must be developed and disseminated.

Another challenge seemed to operate at a cognitive level of information processing and translation between signed language and English. For example, when the hearing signers paused to think about what Keysigns to create, it seemed as though they were first thinking in English because they were muttering to themselves (you can see their lips move in the video). In contrast, when a deaf person was thinking about the Keysigns, their fingers moved instead of their lips. When a hearing person moves their lips, it may indicate that information is being processed from ASL into English before being re-coded into ASL. On the videos, deaf participants can be seen to wiggle or move their fingers indistinctly in thought before responding with clearly signed concepts; an indication that the signed content is not being translated or recoded.

Another unanticipated challenge during this study arose from interpreters’ strong aversion to videotaping their signing. This was the principal reason given by non participants for not contributing Keysigns to the project and for participating interpreters’ incomplete signing. This reaction goes beyond simply disliking being videotaped. Questioning of interpreters revealed that videotaping of their signing was usually conducted to formally assess their abilities during training and certification. Sign language interpreters therefore are very sensitive about videotaping their sign production.

The videotaped Keysigns were analyzed and the sign gloss for each Keysign was recorded. A ‘sign gloss’ is not simply a translation (transliterating) of the signed language into text. Instead, it is an attempt to represent in text form each signed concept along with notations to account for the facial and body grammar that goes with the signs. For the purposes of this paper, these sign glosses have been simplified as text only and are presented in Table 1.

A number of interesting patterns can be seen in the Keysigns contributed by the various participants. Although Keysigns created to describe each session demonstrate various levels of generality and abstraction, for the most part, the Keysigns show a basic level of categorization consistent with Rosch’s (1976) theory of natural categorization. Rosch and others have shown that when people are thinking about the world around them, their preferred level of categorization is at the basic level and they use more basic level terms than subordinate or superordinate ones (Murphy & Wisniewski, 1989; Rosch et al., 1976). Although participants’ Keysigns demonstrate different underlying structures, they are not presented in any obvious hierarchical structure. For example, participants did not describe Presentation 3 as primarily about the broad topic of Science, followed by the more specific domain of Genetics, followed by more specific terms such as structural genetics or genetic mapping. Another interesting pattern can be seen in the use of the Keysign for “Deaf” which is used as a basic level descriptor, as well as a modifier to create subordinate Keysigns (e.g., Inclusion Deaf, Mouse Deaf).
Table 1. Keysigns created by participants

<table>
<thead>
<tr>
<th>Presentation 1</th>
<th>KeySign 1</th>
<th>KeySign 2</th>
<th>KeySign 3</th>
<th>KeySign 4</th>
<th>KeySign 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>educational interpreter</td>
<td>k to 12</td>
<td>university education</td>
<td>children</td>
<td>results</td>
</tr>
<tr>
<td>B</td>
<td>content</td>
<td>skilled</td>
<td>commonalities</td>
<td>education</td>
<td>accessibility</td>
</tr>
<tr>
<td>C</td>
<td>skills learned</td>
<td>education</td>
<td>higher education</td>
<td>access</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>educational interpreting</td>
<td>interpreter quality</td>
<td>university support</td>
<td>interpreter students match</td>
<td>interpreter language acquisition</td>
</tr>
<tr>
<td>E</td>
<td>interpreter</td>
<td>interpreter quality</td>
<td>interpreter education</td>
<td>mentoring</td>
<td>language acquisition</td>
</tr>
<tr>
<td>i</td>
<td>interpreting</td>
<td>deaf</td>
<td>college/university</td>
<td>interpreters skilled</td>
<td>mentoring</td>
</tr>
<tr>
<td>ii</td>
<td>k to 12</td>
<td>deaf</td>
<td>interpretation</td>
<td>education high</td>
<td>stability</td>
</tr>
<tr>
<td>iii</td>
<td>Canadian studies</td>
<td>college/university</td>
<td>interpreters evaluation</td>
<td>inclusion deaf</td>
<td>education</td>
</tr>
<tr>
<td>iv</td>
<td>academic</td>
<td>interpreter</td>
<td>standards</td>
<td>process selective</td>
<td>accommodation</td>
</tr>
<tr>
<td>Presentation 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>academic role</td>
<td>education</td>
<td>service academic</td>
<td>academic</td>
<td>instruction</td>
</tr>
<tr>
<td>B</td>
<td>research</td>
<td>teaching</td>
<td>service</td>
<td>knowledge</td>
<td>content</td>
</tr>
<tr>
<td>D</td>
<td>deaf academic</td>
<td>academic role</td>
<td>working responsibility</td>
<td>committee</td>
<td>participation</td>
</tr>
<tr>
<td>E</td>
<td>role university professor</td>
<td>deaf academic</td>
<td>working responsibility</td>
<td>collaboration</td>
<td>research</td>
</tr>
<tr>
<td>i</td>
<td>deaf academics</td>
<td>teachers</td>
<td>interviewing</td>
<td>people evaluation</td>
<td>children</td>
</tr>
<tr>
<td>ii</td>
<td>management</td>
<td>adult</td>
<td>responsibility</td>
<td>improvement</td>
<td>thesis</td>
</tr>
<tr>
<td>iii</td>
<td>university teaching responsibility</td>
<td>role university</td>
<td>research</td>
<td>advising graduate</td>
<td>curriculum development</td>
</tr>
<tr>
<td>iv</td>
<td>college teaching</td>
<td>teaching workload</td>
<td>research</td>
<td>advising graduate</td>
<td>funding</td>
</tr>
<tr>
<td>Presentation 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>chromosomes</td>
<td>dominant</td>
<td>recessive</td>
<td>chromosomal markers</td>
<td>genetics</td>
</tr>
<tr>
<td>B</td>
<td>genetic hereditary</td>
<td>phenotype</td>
<td>chromosome</td>
<td>recessive</td>
<td>mutations</td>
</tr>
<tr>
<td>C</td>
<td>scientific</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>mouse human genetics same</td>
<td>lab work</td>
<td>mouse analyze</td>
<td>structural genetic mapping</td>
<td>my science signs</td>
</tr>
<tr>
<td>E</td>
<td>genetic hereditary</td>
<td>structural genetic mapping</td>
<td>mouse deaf analyze</td>
<td>signs science</td>
<td>dna same</td>
</tr>
<tr>
<td>i</td>
<td>discuss studies</td>
<td>genetics structure</td>
<td>mouse deaf</td>
<td>mouse circle</td>
<td>phenotype</td>
</tr>
<tr>
<td>ii</td>
<td>rat</td>
<td>phenotype</td>
<td>genotype</td>
<td>structural genetics DNA</td>
<td>sterocilia</td>
</tr>
<tr>
<td>iii</td>
<td>identify</td>
<td>structural genetics</td>
<td>hereditary genetics</td>
<td>hearing</td>
<td>sterocilia</td>
</tr>
<tr>
<td>iv</td>
<td>medical research</td>
<td>DNA</td>
<td>structural genetics</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*A–E — interpreters  i–iv — deaf participants

Conclusion
The writing of this paper presents an example of the unanticipated challenges that arose in the course of this pilot study. In this paper we are discussing visual-spatial language issues using a translated written form that is based on a linear spoken language because
that is the convention for most academic conferences. However, many of the issues under investigation in this study are best expressed and illustrated in sign language and not in a written form. A related challenge was in the elicitation of signed keywords or key concepts from participants. Although the creation of metadata through indexing and categorizing of written language has been a common activity for many years, for sign language users, this is a new way of conceptualizing the language and new concepts for describing it must be developed and disseminated.

Knowledge of the underlying structure of user generated Keysigns can inform the design of new tools for indexing and retrieving sign language content. At the same time, we recognize that Keysigns alone may not perform adequately for all tasks or for all users. As an interim measure we propose that a hybrid approach be adopted where there are text and video-based Keysign indices. Each Keysign will also have a text-based metadata component using a tool such as the EUDICO Linguistic Annotator (ECHO, 2007). The text-based component of the system will include descriptors for each of the five sign characteristics (handshape, palm orientation, movement, location and non-manual signals) that describe the user identified Keysigns. In the next phase of this project, we plan to develop mechanisms for users to create and upload their own Keysigns to tag sign language videos at the alsph.ca website. We also plan to explore the utility of using Keysigns as a mechanism for searching and navigation alone, and in concert with text based metadata.

Acknowledgement
The authors wish to thank Heritage Canada for their support of this project. We also thank Jan Richards, Jim Hardman, Lu Luo and Norma Thompson for all of their suggestions and hard work in organizing the workshop and editing the video material collected. Finally, we acknowledge all of the presenters from the workshop and the participants in the Keysign study who provided their insight and contributions.

References


Visual Knowledge Organization
Towards an International Standard or a Local Institutional Practice?

Abstract
With the aim to digitize visual cultural heritage collections and make them accessible to a broader audience through the Internet, there has been a demand for replacing local knowledge organization (KO) solutions with more international and standardized ones. The argument has been that it is more rational, economical and user friendly; it enhances communication across domains and the user can search for pictures in the same way irrespective of the whereabouts of the local collection. The advantages are obvious. This paper will instead discuss the pitfalls with this development: what might be lost in a process of far-reaching standardization efforts; how local institutional KO solutions might be repressed and knowledge and perspectives invested in those local systems lost.

I wish to make two important points with this paper, firstly that throughout history different institutional settings have developed different ways to collect, manage, and organize pictures. My second point is that this institutional differentiation, when it comes to approaching and describing visual material, has been challenged by digitization. Because digitization creates opportunities for a more global and general access to the material in question, it brings with it certain ambitions to adjust local KO solutions to more standardized KO tools. This paper addresses questions of representation and power; emphasizing how KO tools are embedded with power, and how KO tools and KO practices do not reflect and represent our visual cultural heritage but create it. Empirically the discussion takes its point of departure in a specific, local, institutional setting — The Map and Print Department of The National Library of Sweden. Theoretically the argument is set in a sociocultural, discursive, and domain analytic camp.

Introduction
The digitization of visual cultural heritage collections to make them accessible to a broader audience through the Internet has accentuated the need for more efficient picture indexing tools and retrieval systems. Existing local practices, vocabularies, and indexing tools have been perceived as too specific, exclusive, and too domain-orientated. They have been criticized for not meeting the needs of the general user and limiting the use of the collections (Jörgensen, 1996, 1; 1998, 170–171; 2003, 69–134; Roberts, 1988, 87; Armittage & Enser, 1997, 287, 294; Hollinka et al., 2004, 601, Greenberg, 2001, 921).1

To develop such KO tools there has been a demand for research projects identifying the needs and behaviour of the “average user”.2 Through empirical, cognitive, and perception oriented user studies, researchers are attempting to identify how users in general search for pictures — “in as natural and unconstrained a manner as possible” (Jörgensen, 2003, 6). The effort to meet the demands of the user and making the system user-friendly has been the hallmark of the picture indexing research in the 1990s and 2000s. Researchers within the field of picture indexing have been criticizing former research for being too “system driven”, lacking the “user driven” perspective emanating from the cognitive movement — a perspective that, according to Enser, was introduced very late in the field of picture indexing (Enser, 1995, see also Chen & Rasmussen, 1999, 153; Jörgensen, 2003, 6). Today, as I see it, the cognitive paradigm is very dominating in the field of picture indexing, while more social constructionist ideas are rare. For a further discussion on different paradigms within the field of LIS see Talja et al. (2005).

---

1. This is not a situation specific only in the field of picture indexing; on the contrary, with a more global praxis of information transfer, and a demand for communication over domain and national boundaries, a greater degree of standardization has been necessary in the information transfer field in general. However, although there is a trend in the text oriented KO field in general, toward supporting domain-specific tools and away from cross-domain solutions, this is not yet a very perceptible trend, from what I have seen, within the field of image indexing and retrieval.

2. The effort to meet the demands of the user and making the system user-friendly has been the hallmark of the picture indexing research in the 1990s and 2000s. Researchers within the field of picture indexing have been criticizing former research for being too “system driven”, lacking the “user driven” perspective emanating from the cognitive movement — a perspective that, according to Enser, was introduced very late in the field of picture indexing (Enser, 1995, see also Chen & Rasmussen, 1999, 153; Jörgensen, 2003, 6). Today, as I see it, the cognitive paradigm is very dominating in the field of picture indexing, while more social constructionist ideas are rare. For a further discussion on different paradigms within the field of LIS see Talja et al. (2005).
1996, 2) — and what range of attributes those “general users” tend to use in their search for visual material (Jörgensen, 1998, 164; 2003, 1–7; 203, Choi & Rasmussen, 2003, 499; Hollinka et al., 2004). When identified, the behaviours of, and the attributes asked for by those users, can help to develop more general and standardized tools for picture indexing that can (i) meet the needs of the man on the street, (ii) serve a wide variety of different picture collections, and (iii) enhance cross domain communication (Armitage & Enser, 1997, 287, 294, Jörgensen 1998, 163).

Although this research has advantages — not least in its demand for developing user friendly picture indexing and enhance access to picture collections — this paper criticizes such standardized ambitions based on investigations of how the average user searches for pictures. From a sociocultural and discursive theoretical position, I argue that (i) there is no average user; (ii) information needs and behaviours differ in various institutional settings, (iii) those institutional practices might be a better point of departure for the analysis of picture indexing demands than an analysis of individual users, and (iv) that there is still a need for domain specific solutions.

The sociocultural perspective and the institutional practice

The sociocultural perspective stresses that since the human being is a socially constructed creature, shaped by cultural, linguistic, and material circumstances, no average person — untouched by social conventions and conditions — exists (Hall, 1997; Foucault, 2006; 1991). Different cultural, social, and historical settings, with different tools and material circumstances, will give different experiences and different kinds of knowledge. Stuart Hall describes it as follows: “Knowledge does not operate in a void. It is put to work, through certain technologies and strategies of application, in specific situations, historical contexts and institutional regimes” (Hall, 1997, 49). Consequently, those social settings and tools available will have a decisive impact on how we understand and act in our lives, and, in some circumstance, who we become.

Sociocultural circumstances affect the practices in cultural heritage institutions as well. With different historical traditions, social settings, and material conditions, they will display varying approaches in how to handle, organize, and present cultural items. They will develop different KO solutions; solutions that in the end will frame the way the visual heritage can be presented to, and to some extent understood by, the user. One might say that our cultural heritage is constructed — and not neutrally mediated — through practices and tools used by our cultural heritage institutions.

This sociocultural perspective is far from unknown in the area of Library and Information Science (LIS) and within KO studies. Quite the contrary, in the wake of the linguistic turn within the social sciences and the humanities, many researchers within the field have stressed the importance of paying attention to different cultural and social strategies and solutions in organizing and comprehending knowledge of the world (for instance: Bowker & Leigh Star, 1999; Hjørland, 1997; 2002; Talja et al., 2003; Frohmann, 1994; Olson, 2001). KO systems and classification schemes have been analyzed and their presumptive perspectives have been exposed (Orom, 2003; Hansson, 1999). The consequences — with regard to power — of KO systems favouring some perspectives and disguising others have been analyzed and problematized (Bowker & Leigh Star, 1999; Olson, 2001) and, consequently, KO tools have been regarded as biased and embedded with power (Sundin & Johannisson, 2005).
The library as an institutional setting for picture collections — An example from the National Library of Sweden

In order for me to analyze how picture indexing is affected by social circumstances and institutional practices I chose a specific institutional setting — The Map and Print Department (MPD) of The National Library of Sweden. (For a more comprehensive presentation of this study, see Kjellman, 2006). One reason for this choice was the fact that libraries are not usually associated with the practice of picture management, since books and texts have been the core objects of library commissions. Yet, many libraries have collected and managed picture collections. National libraries in particular can show extremely impressive collections, and The National Library of Sweden is no exception.

So, how has this library specific setting affected the way pictures have been collected, managed and organized? The MPD has actually (i) collected a specific kind of picture material, (ii) developed certain library institution management behaviour towards the material, (iii) and developed specific KO practices. These three aspects have been the focus of my analysis.

The picture material in the collection can be described as being (1) extensive, (2) multifaceted, (3) paper borne, (4) often mass produced and therefore not considered particularly unique or valuable. The library holds millions of objects made in numerous different techniques, with many different functions and with various kinds of motifs. Even material, which from an aesthetic or monetary perspective can be considered not very valuable, has been collected and preserved by the library. One reason for this is of course the law for legal deposits, a unique institutional praxis within the national library setting, which creates conditions for a relatively unreserved form for collecting those pictures that circulate in the day-to-day sphere. It means that the library, to which legal deposits are delivered, operate a unique institutional praxis in the collection of pictures, and that the library, due to that, administers an important and many faceted part of our visual cultural heritage.

Regarding the KO and indexing practices of the material there are at least two aspects worth mentioning here. First: there is a great variation in the cataloguing practices of the material; some collections are very well described in extensive catalogues (especially the older and more unique material) while other collections are treated more rudimentarily. The variation bears witness to the fact that the ordering and indexing tools have evolved historically and pragmatically and thus reflect social and material circumstances, preferences, and focus from different eras. Second: there is in this picture collection a fundamental organizational order which takes its starting point from the identification of people, places, and events; the main categories are portraits, sceneries and historical events. This is an order which in a decisive way distinguishes the library’s indexing procedures and knowledge organizational focal point from, for example, that of museums of various kinds. If the praxis of art museums takes an interest directly in the picture itself as an object, i.e. the material, style, technique, creator of the specific art object; praxis in the library is to orient itself towards that which lies outside the picture, i.e. that which

---

3. Not very surprising since it is texts/books and not visual expressions that have been associated with libraries, science, knowledge, and information traditionally, while pictures have been associated with beauty, entertainment, and emotions. W. J. T. Mitchell shows in his book *Picture Theory* that ever since “Moses denounced the Golden Calf” we have been living in an iconoclasm stressing that it was the spoken or written word that made man into a reflective and rational man, while pictures have been the medium of “the subhuman, the savage, the ‘dumb’ animal, the child, the woman, the masses” (Mitchell, 1994, 2, 24).
the picture refers to, the information or content the picture carries. This is an effect of the institutional practice of the library. The library has, by tradition, focused on text/books and information/knowledge. This orientation has led to a praxis of collecting books and text documents with regard to the content and subjects of the items collected. And a consequence of this focus is that the material and physical aspects of the document are often disregarded, while documenting and informative aspects are emphasized (for a further discussion on libraries’ focus on content and the idea of immateriality of text, documents and information, see Dahlström, 2002, 74; Frohmann, 2004, 6–9; Hayles, 2002, 19).

My investigation showed that the specific institutional setting of this library has developed its own routines and organizational tools emphasizing specific aspects of the visual heritage. However, as mentioned above, when digitizing collections, local KO solutions have been deemed obsolete and in need of replacement, or adjustment, to more national or international standardized solutions. MPD is in this respect no exception. During the last decade, several projects were carried out at the MPD to improve the accessibility of its picture material, not least through digitization and improved KO tools. In those projects there was an ambition to try to achieve broader archive, library, and museum cross domain co-operation by developing shared national vocabularies and indexing tools. The goal has been to develop a joint KO-platform for cultural heritage institutions that provides access to picture material in the same way, with the same KO-tools, irrespective of the whereabouts of the local collection. The development of these projects can be described as a movement from local and internal solutions to a more co-operative and external approach.

Long-term implications of standardization

Standardizing ambitions will, in the long term, have certain consequences. Firstly, different catalogues created within institutional practices are valuable knowledge investments produced and gathered over a long period of time. Not only do they hold valuable information about the items collected, they also bear witness of contemporary interests and preferences; they document the way people once comprehended the material, what they found important to document and what they excluded. A replacement of those catalogues to more standardized solutions will certainly mean an irreparable loss of knowledge. Secondly, a development in that direction will also replace the specific institutional approach — developed within specific settings and invested in organizational praxis — with another perspective representing other preferences and ways of presenting the world. This is the prerequisite for the work of standardization (as the following quotation indicates); the standard you choose will always give priority to someone’s view of the world, and push others into the background: “While unitary documentary languages ensure a maximum of mutual understanding […] they do so by legitimizing a particular ideological and sociopolitical worldview, and by silencing other meanings, voices, and ways of knowing”(Talja et al., 2003, 563). So, the perspective setting the standard will dominate how the visual heritage will be comprehended, and other ways of comprehension will be marginalized. Using the agreed standard will thus limit the ways in which our cultural heritage can be understood and seen.

From this it follows that our KO tools cannot be regarded as neutral — as if they just mediate the world as it is. On the contrary; with these tools we actually create our heritage; with them we construct the present and past of our lives and culture. David Lowenthal, professor of Heritage Studies, gives the following comments on the creative practices of collecting, describing and presenting our cultural heritage:
We still steal, forge, and invent much of our heritage [...] At its best, heritage fabrication is both creative art and act of faith. By means of it we tell ourselves who we are, where we come from, and to what we belong. Ancestral loyalties rest on fraud as well as truth and foment peril along with pride. We cannot escape dependency on this motley and peccable heritage. But we can learn to face its fictions and forgive its flaws as integral to its strengths. (Lowenthal, 1998, xvii)

As Lowenthal stresses, we cannot choose not to do this, i.e. using tools and practises that bias the representation of our cultural heritage. What we can do is to become aware of our actions and the effects of our chosen tools. A development that enhances more global information and communication strategy with the help of more standardised tools, not bound to domain or nation specific solutions, will in this respect, put local variations at risk. With an example from the tourism and museum industry, Lowenthal argues that we have a development that standardizes — in a Western way — how the cultural heritage from different areas is being exposed:

Most heritage is amassed by particular groups, but media diffusion and global networks make these hoards ever more common coin. [...] Display and tourism layer diverse legacies with common facades. The same multinationals finance restoration in Prague and Peru, using techniques devised in Rome and London. Legacies of nature, prehistory, art, and architecture are hyped in terms ever more alike. [...] The language of heritage that suffuses the world is mainly Western. (Lowenthal, 1998, 5)

Conclusions
Although standardization benefits from rationality and stability — and maximizes mutual understanding — there are negative outcomes as well. In contrast to homogenizing ambitions, I argue for the elaboration of differentiated KO solutions that are sensitive to domain specific demands. During recent years several researchers within the field of LIS, who from experience of standardization of vocabulary in the form of classification systems and other nomenclature always tend to emphasize certain perspectives at the cost of others, have argued for alternative systems based on multi-use, flexibility and openness (Bowker & Leigh Star, 1999; Olson, 2001). Even if this results in a lack of rationality, order and stability, it creates, on the one hand, possibilities for making otherwise marginalized perspectives visible and, on the other hand, an understanding of the constructional nature of knowledge. It is actually in the meeting between different information systems that this constructional ability can be revealed. Bowker & Leigh Star write: “As the information systems of the world expand and flow into each other, and more kinds of people use them for more different things, it becomes harder to hold to pure or universal ideas about representation or information” (Bowker & Leigh Star, 1999, 301). Consequently, the encounter between various KO solutions tends to expose the constructional effects of knowledge representations, while a far developed standardization will run the risk of presenting the world as something stable and universal. Another consequence of this perspective is that our cultural heritage — visual and textual — can no longer be viewed as something that simply is. Instead, it must be seen as constructed through our language, our institutional setting, and our tools. When constructing languages, institutional practices, and KO tools of different kinds, we also construct our heritage. Through these actions we are responsible for whose perspective will be seen, and whose will not.

References


Aaron Loehrlein — Indiana University

The Benefits of Participating in a Form of Life: Interpretations of Complex Concepts among Experts and Novices in Records Management

Abstract
This study is an empirical investigation into the specific advantages gained through familiarity with forms of life and their accompanying language games, as understood by Ludwig Wittgenstein (1953), when interpreting and conceptualizing about complex concepts that are represented within that form of life. Subjects with experience in a specific form of life, e.g., records management, were presented with passages representing complex concepts embedded in that form life. The subjects were asked to rank interpretations of those passages in order to determine which aspects of the complex concepts they found most salient. Their responses were compared with those of two groups of novices in records management. The differences in responses between the three groups highlight the specific knowledge effects at work. Finally, the paper addresses methods for distinguishing between familiarity with concepts and familiarity with language.

Introduction
In his Philosophical Investigations (1953), Ludwig Wittgenstein battles against “the bewitchment of our intelligence by the means of language” (aphorism 109). Instead of focusing solely on the meanings of words, he exhorts us to consider the uses of words and the situations in which those uses occur. He refers to these concepts as language games and forms of life. Although he offers no explicit definitions for either concept, he maintains that language games provide the framework in which an example of the word’s usage “makes sense” (Blair, 1990). As such, they are the means by which one learns new words, or new uses for familiar words, within the context of a particular community or culture. Language games do not provide sets of rules for appropriate speech or actions, but a natural structure that encourages a certain manner of speech or action (Blair, 1990). Wittgenstein refers to these structures as games in order to emphasize that the speaking of language is part of an activity, also known as a form of life (Wittgenstein, 1953). For Wittgenstein, forms of life include such activities as giving and obeying orders, describing the appearance of an object, reporting an event, forming and testing hypotheses, or even guessing riddles and telling jokes (ibid.). One cannot gain mastery of the language employed by a community without first understanding the forms of life in which those phrases are used (Wittgenstein, 1966). Furthermore, an understanding of a form of life can only be gained though active participation in that form of life (Blair, 1990).

The profession of records management entails many such forms of life. Records managers are concerned with the analysis of how information can best be used as “a tool within a particular organization” (Choksy, 2006, xviii). Some of their principle activities include determining what records should be created for a given business process, ensuring that the records are maintained in safe environments and that they are disposed of properly when no longer needed or require. Each of these actions must be also in compliance with legal and regulatory requirements (Brumm & Carlisle, 2005). These forms of life naturally entail language games that provide a structure for how language is used. The papers and articles that form the records management discourse often do not include words or phrases that would be unfamiliar to those outside the profession. In addition, most of the key terms of records management, such as record, retention, and compliance, are used in a manner that is broadly consistent with their usage among the general
population. Nevertheless, the language used within this discourse might follow subtle constraints, so that the main point or purpose of a given passage would not be fully understood by non-records managers.

This study is an investigation into the specific effects associated with this phenomenon. Although records management is the object of study, the purpose is not to examine records management per se. Rather it is designed to identify the processes associated with knowledge effects and to determine the extent to which those processes are associated with participation in a form of life, as opposed to the direct application of knowledge. The study examines complex concepts that are not established enough to be represented by a single term or a short phrase, but require more lengthy description such as a passage of text. Because concepts of this sort do not tend to have linguistic labels, they are likely to be referenced primarily in the context of a specific activity or form life (e.g., retaining records, etc.). Therefore, participation in a form of life might have an effect on which aspects of the complex concept are considered most salient. By shedding light on such processes, this study endeavors to examine some of the problems that might occur when multiple groups of people, each with their own base of knowledge and experience, attempt to communicate with each other or attempt to use a common controlled vocabulary.

In an earlier investigation into the organization of complex concepts, Chi, Feltovich, and Glaser (1981) presented a series of physics problems to a group of physics experts and a group of novices and asked the subjects to categorize each problem. Chi et al. found the experts categorized the problems according to general physics principles, while the novices tended to focus on the surface similarity of the problems (e.g., problems involving springs or inclined planes). In addition, Chi et al. and Chi, Hutchison, and Robin (1989) found that, when people categorized items in a domain for which they had expert knowledge, they were effective at applying their knowledge to the categorization. A series of studies conducted by Douglas Medin (e.g., Medin, Lynch, Coley, & Atran, 1997) found that experts and novices attend to different properties when deciding if an item is a good example of a given category. In addition, Barsalou (1982) found that the salience of certain properties of objects can depend on the circumstances in which they were used. For example, the property floats might not ordinarily be salient of the entity ball, even though it is true of most balls. However, the salience of that property is likely to increase if one is in need of a floatation device.

Concepts that are represented by convenient linguistic labels, such as tree, are likely to be well-established with readily apparent properties such as height, and members, such as elm or even tree in my backyard. In their study of the effects of expertise on one’s conception of tree, Lynch, Coley, and Medin (2000) provided subjects with specific types of tree and asked the subjects to rate each type in terms of certain preselected properties, including height and weediness [sic]. The subjects also rated how well each type of tree fit the category of trees in general. They found that subjects with expert knowledge tended to rate the tallest trees as most the most typical, while novices rated the trees that they were personally most familiar with as the most typical. Studies such as Lynch et al. generally do not ask the subjects to provide descriptions of the concepts, since such descriptions are likely to represent the subject’s understanding of the accepted meaning of the terms used to refer to the concepts. Instead, they present the subject with specific tasks in order to determine the probable cause for different behaviors in performance of the tasks.

In contrast, concepts that have no linguistic label and that can only be represented with a complex description are unlikely to have readily apparent properties or exemplars.
Therefore, when studying these concepts, it is not practical to measure the difference in how they used in terms of the salience of its properties or exemplars. Instead, subjects in the current study were provided with a set of interpretations of a passage and were asked to rank them in terms of their appropriateness to the passage (see Table 1). Some interpretations (“Main”) are designed to represent one of the main points of the passage, other interpretations (“Not Main”) represent points made by the passage other than the main points, while additional interpretations (“False”) represent points not made by the passage. In addition, some interpretations use similar wording to the passage, while other interpretations seek to convey the same concept but with dissimilar wording. Because subjects were asked to rank the interpretations, they did not need to determine which interpretations are valid for the passage and which are not, but simply sequenced them by their perceived appropriateness. In addition, the subjects could not refer back to the passage when ranking the interpretations, which forced them to rely on their intuitions and their short-term memory.

Table 1. Sample passage and interpretations

<table>
<thead>
<tr>
<th>Passage</th>
<th>Interpretations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Record retention policies and procedures must be regularly and fully</td>
<td>Similar</td>
</tr>
<tr>
<td>implemented by corporate business units. If this is not done, preparing</td>
<td>To be legally acceptable, record retention policies must be not only developed,</td>
</tr>
<tr>
<td>retention guidelines is merely a time-consuming exercise. For corporate</td>
<td>but implemented.</td>
</tr>
<tr>
<td>retention practices to be considered legally acceptable, records must</td>
<td>Dissimilar</td>
</tr>
<tr>
<td>be discarded in the normal course of business when indicated in</td>
<td>A plan for ultimate disposition of records will provide no benefit if it is not</td>
</tr>
<tr>
<td>retention policies and procedures.¹</td>
<td>acted upon.</td>
</tr>
</tbody>
</table>

Main
Preparation guidelines and discarding records can both be time-consuming  |
exercises. Successful corporations must ensure that the retention and     |
disposal of corporate records is given sufficient resources.               |

Not Main
Legally acceptable retention guidelines require that records be retained  |
in the normal course of business. Legal policies require corporations to   |
determine the records that they have discarded.                            |

False

Subjects, Materials and Procedure
A total of 29 subjects were recruited for this study: ten professional records managers, nine master’s degree students in library and information science (LIS) and ten undergraduate students. None of the LIS graduate students or the undergraduate students had any knowledge of or formal education in archives or records management. The subjects were each offered a $10 gift card for participating in the study.

The stimuli consisted of ten passages of one to two paragraphs in length taken from larger article or paper. Five of the passages were taken from records management literature (e.g., *Information Management Journal*). The remaining five passages were taken from *The New Yorker*, a general interest periodical. None of the passages were less than two years old, reducing the chance that a subject was already familiar with the passage. For each passage, six interpretations were provided as described above. Three professional records managers acted as judges, reviewing and validating the interpretations of

records management passages. In the rare cases where the judges disagreed, the change suggested by the majority was adopted. A separate group of two Ph.D. students reviewed and validated the interpretations of general interest passages. Finally, two additional LIS Ph.D. students served as subjects for a pre-test of the questionnaire. In the pre-test, fatigue seemed to have a strong effect. Therefore, the number of trials was minimized and practice trials were eliminated.

For each trial, the subject read the passage on a computer screen. Once the subject was ready she pressed a button to replace the passages with six interpretations of the passage. She then ranked the interpretations according to their appropriateness to the passage. Subjects were assigned to one of three different random orders of the trials to control for carryover effects. Subjects took an average of 25 minutes to complete the ten trials.

In the records management literature, records managers are considered experts, while the other two groups are considered novices. Because records managers and graduate students are usually college graduates, they are assumed to have reasonably high standards of literacy, which undergraduates may or may not share. Therefore, in the general literature, both graduate students and records managers are considered experts, while undergraduates are considered novices. The first hypothesis (H₁) is that experts will be able to consistently distinguish between the Main, Not Main, and False interpretations. The second hypothesis (H₂) is that novices will be able to distinguish between the interpretations that are valid and invalid, but will be less effective at identifying the interpretations that represent the main point of the passage. The third hypothesis (H₃) is that novices are likely to have less confidence with the concepts represented in the passage. Therefore, they will be more likely to give high rankings to interpretations that use wording similar to the passage. In contrast, experts are likely to pay attention to deeper similarities between the interpretations and the passage. Therefore, the fourth hypothesis (H₄) is that the rankings given by experts will not significantly differ between the Similar and Dissimilar conditions.

Results and Discussion
For the records management passages (see Figure 1), all three groups of subjects gave significantly lower rankings to False interpretations than to Main interpretations. Records managers gave significantly higher rankings to Main interpretations (M = 4.77) than to Not Main interpretations (M = 3.18) t(198) = 7.66, p < .01, as did LIS graduate students (M₁ = 4.39, M₂ = 3.3) t(178) = 4.53, p < .01. The rankings given by undergraduate students did not significantly differ between the Main and Not Main condition. Therefore, the data concerning the records management literature supported H₁, while H₂ was supported for undergraduates but not for LIS graduate students.

For the general interest passages (see Figure 2), all three groups gave significantly higher rankings to Main interpretations than to Not Main or False interpretations. Records managers gave significantly higher rankings to Not Main passages (M = 3.20) than to False passages (M = 2.53) t(198) = 3.30, p < .01. However, the rankings provided by graduate students and undergraduates did not significantly differ between the Not Main and False conditions. Because the LIS graduate students were considered experts in regards to the general interest literature, H₁ is supported only for the records managers. In

2. Because the rankings are not completely independent of each other, the t tests described in this study provide an indication of general trends, but do not conclusively demonstrate significant effects or the lack thereof.
addition, the two other groups had trouble distinguishing between the Not Main and False conditions as opposed to the Main and Not Main conditions as predicted by H2. On the whole, the LIS graduate students behaved like experts in the records management literature, which they had no exposure with, but behaved like novices with the general interest literature. It is possible that the background and experience of the LIS graduate students was not in line with the assumptions of the study, i.e., it is possible that many had a higher exposure to academic writing that was similar to records management literature than they had to general interest periodicals. In addition, when a group of subjects had difficulty making a distinction between the types of interpretations, the problem occurred between the Main and Not Main interpretations for the records management literature and between the Not Main and False interpretations for the general interest literature. This suggests that the False interpretations were more salient in the specialized literature, but the Main interpretations were more salient in the general literature.

In the records management literature (Figure 3), records managers provided significantly higher rankings for Main Similar interpretations than Main Dissimilar interpretations ($M_1 = 5.7, M_2 = 3.84$) $t(98) = 8.72, p < .01$ and significantly higher rankings for Not Main/Similar than Not Main/Dissimilar interpretations ($M_1 = 3.44, M_2 = 2.92$) $t(98) = 1.72, p < .05$. LIS graduate students provided significantly higher rankings for Main/Similar interpretations than Main/Dissimilar interpretations ($M_1 = 5.20, M_2 = 3.58$) $t(88)$
300

\[ t(98) = 1.93, p < .05 \]

No other significant differences were found between the Similar and Dissimilar conditions, including any of the results from the general interest literature (see Figure 4). Therefore, neither \( H_3 \) nor \( H_4 \) are supported by the data. One possible explanation is that passages with similar wording are more likely to represent the precise concept expressed in the passage. This might indicate that the terms used in the passage are part of a controlled vocabulary that is used by people who participate in records management activities. If that is the case, then the similarly-worded interpretations might match the specific purpose of the passage to a greater extent than would be appreciated by novices.

One question yet to be addressed is whether the differences in performance between experts and novices is due to participation in activities that make up a form of life, or due to the direct application of specific knowledge that is relevant to the topic. In other words, do the experts behave differently because they have a sense for the “natural structure” of records management activities, which encourages a certain manner of speech and action, or is the difference in behavior simply due to the fact that the experts possess more factual knowledge? The responses to one of the records management passages might shed light on this question. That passage contains statements that were deemed inaccurate by the records management judges. The judges felt that one interpretation represented the main point of the article, but made an inaccurate statement about records management in general. If the records managers were simply applying their knowledge of records management to the interpretation, it seems likely that they would give a relative low ranking to an interpretation that they felt was inaccurate, regardless of whether it was appropriate for the passage. Alternately, if participation in records management as a form of life gives one greater mastery of the language used in that form of life, then it seems likely that records managers would more easily grasp the purpose of the passage and identify the interpretations that are appropriate to that purpose, whether they feel that purpose is valid or not.

Both records managers and graduate students gave significantly lower rankings to the Main interpretation that was considered invalid for records management in general than to the other Main interpretation of that passage. Although the two groups of subjects did not significantly differ from each other, the variance for the LIS graduate students was higher, suggesting that there was more diversity of opinion among that group than among the records managers. These results suggests that the records managers’ higher levels of knowledge were working against them in this case, since the fact that the interpretation was invalid overshadowed the fact that it was an appropriate representation of the passage. However, the fact that this analysis is post hoc and has a low sample size makes it difficult to form conclusions from the data.

**Conclusion**

This study has examined the specific effects that participation in a form of life has on the apprehension of textual representations of complex, context-dependent concepts that find utility within that form of life. It has found that specific wording has a great effect on the use of these complex concepts. Changes to wording seem to result in a great deal less confidence among experts as to whether the original purpose of the statement remains intact, even if the change to wording is deemed to be equivalent by expert judges. The results also seem to suggest that the salience of either the main purpose of the passage
or statements that are valid for the passage tend to be lower for novices than for experts. Further study is needed to determine the specific conditions that are most likely to lend themselves to specific modifications of salience on the part of novices. Finally, preliminary results suggest that experts are more inclined to apply their knowledge directly than they are to make use of the language games that they use. Future studies will compare experts’ ability to interpret passages that use familiar concepts with their ability to interpret passages that use familiar language.

References
Discourse Community Analysis
Sense Construction Versus Non-Sense Construction

Abstract
This paper examines the nature of the political discourse of international organizations (IMF, world Bank, the UN, European Union, African Union, The Arab League, etc.). The study originates from a fundamental paradox: how can we use the same descriptive linguistic tools which we use in analyzing the production of sense for the production of non-sense characterizing this type of discourse? The corpus analyzed is a collection of texts produced the international community on the Darfur political and humanitarian crisis. The high frequency of words and terms occurring in this type of production contributes in losing their meaning which is contrary to the appreciation of terms occurring in an expert discourse for information retrieval or lexical resources purposes where the high frequency of a term consolidates its belonging to a specific field of knowledge. How can this paradox be explained?

1. Introduction
The term discourse community links the terms ‘discourse’, a concept describing all forms of communication that contribute to a particular, institutionalized way of thinking; and ‘community’, which in this case refers to the people who use, and therefore help to create a particular discourse. “Producing texts within a discourse community,” according to Patricia Bizzell, “cannot take place unless the writer can define his goals in terms of the community’s interpretive conventions.” (Bizzel, 1992). In other words, one cannot simply produce any text. It must fit the standards of the discourse community to which it is appealing. If one wants to become a member of a certain discourse community, it requires more than learning the lingo. It requires understanding concepts and expectations set up within that community.

Discourse community analysis is thus a type of discourse analysis which can be defined as analyzing written, spoken or signed language use. The objects of discourse analysis (discourse, writing, talk, conversation communicative events, etc) are variously defined in terms of coherent sequences of sentences, propositions. Contrary to much of traditional linguistics, discourse analysts not only study language use ‘beyond the sentence boundary’, but also prefer to analyze ‘naturally occurring’ language use, and not invented examples.

Discourse analysis has been taken up in a variety of social science disciplines, including linguistics, anthropology, sociology cognitive psychology, international relations and communication studies, each of which is subject to its own assumptions, dimensions of analysis, and methodologies. Topics of interest to discourse analysts consists in the various levels or dimensions of discourse, including significant body language such as sounds, intonations, gestures, the discourse itself (syntax, the lexicon, style, rhetoric, meaning, strategies and other aspects of interaction) and the genres of discourse (various types of discourse in politics, the media, education, science, business, etc.); It also looks at the relations between discourse and the emergence of sentence syntax; the relations between text (discourse) and context; the relations between discourse and power; the relations between discourse and interaction; the relations between discourse, cognition and memory.

This paper originates from a fundamental paradox: how can we use the same descriptive linguistic tools which we use in analyzing the production of sense for the production
of non-sense? My work essentially deals with building linguistic resources for information systems (glossaries, terminologies, ontologies etc.). This process involves collecting, analyzing and describing the linguistic elements from corpora. Discourse analysis although uses the same description tools involves also a ‘beyond the sentence boundary’ analysis. The corpus analyzed is a collection of texts produced by the international community. The terms, understood in a terminological sense, occurring in this type of discourse lose their sense. In other words, one can understand their normal/conventional sense only outside of the context (if they are taken separately). How can this paradox be explained? Their high frequency contributes in losing their sense which is contrary to the appreciation of terms occurring in an expert discourse (scientific and technical discourse.) or in information retrieval where the high frequency of a term consolidates its belonging to a specific field of knowledge. In other words, the terms are used so much that their specific meaning fades away, that is, the terms become empty symbols. Reality description is essential in the case of expert discourse, refereeing to specific concept and their interrelation. This is true of scientific and technical discourse. As for the political discourse of international organizations it seems disconnected from the reality it is referring to unlike the expert discourse mentioned above in which not only terms should adhere to concepts but in addition their standardization is an essential activity.

2. Discourse analysis: sense construction vs non-sense construction

One of the purposes of information and communication sciences is the construction and sharing of meaning. This process involves building reference tools such as thesauri for indexing and retrieving an increasing digital data. Meaning construction through these tools is based on well known conceptual methods. To guarantee communication between the various actors of a field of knowledge, or what we can call a ‘community of experts’ which uses its own discourse these tools should be rigorously and coherently designed. The term holds a univocal relationship with the concept hence normalizing terminologies. Normalization is an activity characterizing many organizations producing and validating such specific terminologies (such as la Délégation de la langue Française in the French context for instance). These coherent approaches are in total contradiction with the implicit meaning construction characterizing the political discourse of international organizations, a rather stilted and ‘politically correct’ form of logorrhea. A recent study (Rist, 200) shows by a variety of examples how this type of discourse is characterized by its non-sense. Thus the international community could be said to use an approach diametrically opposed to the one I have been using so far. This paradox is particularly interesting and calls for a thorough examination of information production contexts. Meaning should normally be contextualized that is why it would be interesting to study the following aspects: who are the “producers” of the discourse. Who are the recipients/targets? What kind of impact does it have on the recipients/targets? The conceptual network centred on a scientific or technical term should help putting the term into a coherent conceptual environment which will help understanding its scope and meaning. If we look now to the international organizations discourse, also called “international rhetoric”, (Rist, 2002) we find a peculiar conceptual network which strengthens the non-sense in a ‘coherent’ manner.

If we examine the political discourse of international organizations we can notice that it is largely symbolical and that its operating field is in itself symbolical. Reality is manipulated by the discourse in order to conform it to the symbolical contents that have been pre-selected. Distortions are denied because behind this symbolism contains a solid ideology
all the stronger for being implicit. Since the end of communism a logico-semantic setting was born and a number of terms like empowerment, stakeholders, transparency, good governance, level playing field, human rights, free market, gender, should be understood, not in their univocal original meaning but rather as symbols. In a way this harks back to the old Communist discourse in which words such as proletariat, working class, masses progress, popular struggle and so on had a “coded” meaning which had only a distant relation either to the political science vocabulary or even to the original Marxist vocabulary it was supposed to have been borrowed from. In analysing the corpus related to the Darfur crisis, a specific discourse genre, I will refer to the studies undertaken by Maingueneau (2002) and Krieg-Planque (2007) which I consider as particularly relevant.

The impact of international rhetoric on its audience can be compared to the Indian Chief discourse described in Pierre Clastre (1974):

Parler est pour le chef une obligation impérative, la tribu veut l'entendre : un chef silencieux n’est plus un chef [...] Ce n’est pas de l’esthétique qu’il est question mais de politique. La parole du chef n’est pas dite pour être écoutée. Parce que littéralement, le chef ne dit, fort prolixement, rien. Son discours consiste, pour l’essentiel, en une célébration, maintes fois répétées, des normes de vie traditionnelles.

(Clastres, 1974)

We translated this quotation as follows: For the Chief, speaking is an absolute imperative. The tribe wants to hear him and a silent chief is not a chief any more. This has nothing to do with aesthetics, it is essentially political. The speech issuing from the Chief’s lips is not made to be listened to. Because, if we consider his discourse in depth, the Chief, although he speaks a lot, says literally nothing. His discourse is essentially a frequently reiterated celebration of the norms of [the tribe’s] social life.

3. Analyzing texts related to Darfur crisis corpus

The corpus we are analyzing is representative of a discourse community. We chose texts related to the Darfur political and humanitarian crisis as we mentioned above. The reason of our choice is our knowledge of the context which is essential for analyzing the corpus. All the texts are related to a single subject: Conflict Early Warning by various international entities. We are giving hereafter some examples, i.e. texts showing the nonsense followed by our comments.

Example 1. “By presenting a credible military threat a third party force seeks to convince all conflictual parties that violence will not succeed. International force is brought to bear not to defeat but to neutralize the local forces and to reduce the expected gains of continued fighting. The political objective is to become a successful arbiter of disputes and to persuade the contending parties they have no alternative but to negotiate. The point is to deny victory to one side in order to create a military stalemate” [David Carment & Dane Rowland: Explaining multilateral intervention in ethnic conflicts: a Game theoretic Approach. ISA Paper. San Diego. 1996.]

Absurd of course since one does not see how one side would be deterred without fighting and even less persuaded that it has no alternative but to negotiate. The author was a game theorist writing on Congo-Zaire. He did not have ANY ground experience.

Example 2. “The most ambitious and focused early warning system yet established in the UN is the Humanitarian Early Warning System (HEWS) of the Department of Human Affairs (DHA). It involves a number of professional staff and a significant computer
capacity. It incorporates a multitude of sources (statistical and textual) to allow monitoring in over 100 nations. However, the system has yet to produce a single “early warning” since it has been in operation since July 1995. (...). The solution to this political problem is to develop a new norm for early warning in which the UN Secretary General draws attention to potential conflict at an early stage in spite of the reluctance of the members.”[pp. 168–169]A Walter Dorn: “Early (and late) warning by the UN Secretary General Synergy in Early Warning. University of York Toronto 15–18 Mar. 1997].

This text is a model of nonsense. The author first explains how a certain system is designed “to allow monitoring of 100 nations”. And he then explains how it does not work at all, since it has “not produced a single early warning” since it has been in operation since July 1995. Today, thirteen years later, it still has not produced anything!

Example 3. “As the 1996 joint donor evaluation of the Rwanda conflict found, what is needed is not so much information but policy-oriented analysis that will suggest logical operational responses. The debate on early warning has not yet moved forward to deal with the issue of the process link between early warning analysis and effective preventive action. This may be because early warning praxis is not effective in producing analysis that clearly present options for effective preventive action and rapid engagement policy.” John G. Cockell (Department of Foreign Affairs. Government of Canada, collected by the University de York): “Towards Response-oriented Early Warning analysis: Policy and Operational Considerations”.

Meaningless because what it says (once you pierce the language barrier) is that if only “early warning” provided guidelines for preventive action this would work. This is in complete neglect of reality. “Preventive action” (which has never occurred) is linked with political choices, not with “analysis of policy options”. In other words this text says that if a “response-oriented” early warning system existed it would be enough to produce adequate response to violent crisis. There is no any reference to the reality.

4. Discussion
A discourse community has several aspects beyond its technical linguistic aspect. One is the reinforcement of the group’s norms, meaning that a minimum median line has to be adhered to. If a speaker strays too far from this norm his first punishment will be not to be taken seriously. If he strays too far, the second and more serious punishment will lead him to be deprived from his job or at least to see his work perspectives reduced.

But an interesting point is that such “discourse communities” cannot at the same time simply fold back into an autistic reality. They have to deal with the real world which forces them to admit (even if only at the margins) forms of discursive heterodoxy in order to refresh the contents of the discourse and to keep it somewhat connected to reality. A good example of the failure to integrate practically such discursive heterodoxy is the failure of the Communist system between 1965 and 1989 to renew its discourse on the surrounding world, leading to an increased disconnection between the accepted discourse and what it was referring to.

5. Types of tools needed for analysing and extracting terms and collocations
In our previous researches we largely dealt with term extraction tools, their description and their evaluation. We don’t think that these tools will be sufficient for analyzing this type of discourse even if they can be partially relevant, i.e. extracting nominal construc-
tions. For discourse analysis we need more holistic tools capable of analyzing the sentences and find collocations. The analysis should involve three essential aspects are i) the relative authorship weight of the text producer since the impact of the nonsense is greater if the text producer is “important”; ii) the reality about which the discourse is produced and; iii) the impact of the discourse on the audience (see also above). These three aspects involve efficient semantic and pragmatic analyzers and we are not really sure that this type of this discourse analysis would be possible to automate. Moreover, there is a pre-established potential for receptivity to different types of discourses depending on the nature of the audience.

6. Conclusion
We assume that the human involvement in analyzing this type of discourse is crucial since existing tools are still very primitive as far as semantic and pragmatic analyses are concerned. Their use can be helpful in identifying terms and collocations and their contexts but they will be insufficient for analyzing the sense or non-sense characterizing this type of discourse.

References
The Impact of the Sloan Digital Sky Survey on Astronomical Research
The Role of Culture, Identity, and International Collaboration

Abstract
We investigate the influence of culture and identity (geographic location) on the constitution of a specific research field. Using as case study the Sloan Digital Sky Survey (SDSS) project in the Astronomy field, we analyzed texts from bibliographic records of publications along three cultural and geographic axes: US only publications, non-US publications and international collaboration. Using three text mining systems (CiteSpace, TermWatch and PEx), we were able to automatically identify the topics specific to each cultural and geographic region as well as isolate the core research topics common to all geographic zones. The results tended to show that US-only and non-US research in this field shared more commonalities with international collaboration than with one another, thus indicating that the former two (US-only and non-US) research focused on rather distinct topics.

1. Introduction
Culture and identity play a major role in the complex processes involved in knowledge creation, representation and acquisition. However, these two parameters have rarely been the focus of automated methods for knowledge representation. We aim to investigate if culture and identity (geographic location) influence the development of a specific research field. We take as case study the Sloan Digital Sky Survey (SDSS) project in the field of Astronomy. SDSS is funded by NASA and the National Science Foundation in the US and aims to collect high quality data for astronomical research. The availability of this data has led to an increasing number of discoveries in astronomical research. Given that this project is funded and operated in the U.S., a natural question would be whether the research themes undertaken by astronomers in the U.S. differ significantly from their counterparts in other countries and regions such as Europe and Asia.

2. Methodology
Our data consisted of a total of 1456 bibliographic records, retrieved from the Web of Science database using a query containing the keywords “SDSS” or “SDSS Digit*”. These records covered the period between 1998–2007. Among the 1456 publications, 379 were made by US institutions only, 459 by non-US institutions, and 618 are joint publications between US institutions and non-US institutions. We call this 3rd set International collaboration. We used three text mining systems to perform our analysis: CiteSpace (Chen 2006), Projection Explorer (Lopes et al., 2007) and TermWatch (SanJuan & Ibekwe-SanJuan 2006). These systems sought to highlight four aspects of the study: geo-spatial mapping; detection of salient topics; mining for association rules and finally a comparative analysis of the topics from each discourse community.

3. Geo-spatial mapping of SDSS authors
A first level of macroscopic analysis is to visualize the geo-spatial distribution of publications across the world. This was done using Google Earth and CiteSpace (Chen 2006). Owing to reasons of space, we show maps for the US region only.

Figure 1. Geospatial map of collaboration in the US-only publications

4. Structure of SDSS research by cultural and geographic regions
Salient topics were identified by applying natural language processing and information extraction techniques to SDSS-related publications. Salient topics are represented in terms of author-defined keywords, noun phrases extracted from the title and abstract fields of each record by CiteSpace. TermWatch was then used to obtain a global view of research topics in the three data sets based on terminological variations.

4.1 Research topics structure in the US-only institutions
375 publications were made by US-only authors in our data set. At the term variation level, a map of research topics was obtained with TermWatch (figure 2). Green nodes denote more recent topics (2005–2007). Pink color are topics whose terms appeared between 2002–2004. Shades of blue indicate topics found in earlier publications made between 1996–1998 (light blue) and 1999–2001 (deep blue). TermWatch identified central and peripheral atoms using the graph decomposition algorithm described in Biha et al., (2007). The most central cluster labeled “halo mass function” seems to be focused on galaxy clustering and formation models. On the whole, the majority of the topics appeared in the most recent period of the corpus (2005–2007).

CiteSpace was used to obtain association networks from titles and abstracts fields of the publications in this dataset. For reasons of space, we cannot show the network obtained from the US-only publications. However, the first cluster includes topics related to “black holes” (bh) such as “velocity dispersion, local bh mass density, bh mass, bh merger, cluster galaxy evolution”. A second middle cluster focused on star formation, including terms like “poststarburst galaxies, emission line, strong balmer absorption line”. A third cluster dealt with “galaxy formation model”, including terms like “quasar luminosity function, halo mass, satellite galaxies, dark matter halo, host galaxies”. The two systems, CiteSpace and TermWatch highlighted some common or semantically close terms even though they used different techniques to extract the terms. The following terms in CiteSpace’s association networks “cdm model, velocity dispersion, velocity distribution, cluster galaxy evolution, star formation” appeared either in the exact form or as semantic variants in TermWatch’s central atom (cold dark matter model, high velocity, cluster galaxy, star formation rate).
4.2. Structure of SDSS research outside the US

Four hundred and fifty-nine publications were made by non-US authors. The titles and abstract fields were analyzed by TermWatch. The map obtained showed that there is no one central atom in contrast to the US-only research. SDSS research outside the US seem to be articulated around five major topics labeled “supernova type ia, star formation rate, black hole, syfert galaxy and nearby cluster”. To gain further insight into the particularities of research on SDSS outside the US, a country-by-country analysis was performed by CiteSpace. We show results for two other prominent countries: UK and China. SDSS research in the UK seem to be characterized by the terms “CDM cosmology (CDM = cold dark matter), sdss spectra and wavelength range. SDSS research in China contains terms such as sdss spectrum, sdss dr1 (data release 1) and stellar velocity dispersion. The cluster on the right deals with topics such as double-peaked broad emission line, dimensionless accretion rate, black hole mass, eddington ratio and oxygen abundance.

Figure 3. Clusters derived from papers by institutions in the U.K ($p = .05$).
4.3. Structure of international collaboration

This third set are publications made simultaneously by US and non-US authors. 618 records were concerned. The following map obtained by TermWatch shows the layout of research topics (figure 5). Like the non-US research, international collaboration on SDSS is not articulated around a unique center. Several subgroups of research topics are connected through chains of intermediary topics. We have circled and labeled the cluster at the center of the different subgroups: cosmological parameter, galaxy-galaxy weak lensing, faint end slope, sagittarius dwarf, stellar mass, fifth lensed image, complete gunn-peterson trough.

5. Comparative analysis of topics across geographical and cultural regions

Here, we seek to determine if there is a core set of research concerns shared by authors regardless of geographical origins, across the three data sets. To this end, we utilized the two systems PEx and TermWatch to perform a more detailed analysis.
5.1 Association rules derived from the SDSS literature by geographic regions

Association rules (ARs) are implications extracted from transaction databases. In the framework of automatic discourse analysis, ARs can be used to detect implications between domain terms supported by their co-occurrences. For instance, an association rule might find that the word “matter” always implies “halo” because the two co-occur more often than not. This is indeed supported by eight documents in the corpus. Locally Weighted Association Rules (LWR) (Lopes et al., 2007) were preliminarily used to identify salient topics from the three data sets. To selectively extract rules, LWR gives more weight to association rules in which words are specific (local) to a subset of documents from the corpus. Table 1 shows the intersections in terms of ARs found in the three data sets. The first row contains terms found uniquely on one of the sets of rules. Terms on the other rows were found on two or more of the datasets. Some remarks can be made from the list produced by the ARs: (i) as the algorithm privileges words that are specific to each subset of the corpus, it is no surprise that there are few common terms; (ii) Cosmic Microwave Background (CMB) is found to be common to both US and non-US, however WMAP (a probe) which detects CMD is found only in the collaboration subset, thus indicating a more widespread subject; (iii) studies regarding quasars, halos and satellites seem to be particular to US-only research; (iv) studies using or related to photometry (photometric, photometry, imaging, luminosity) are the prominently found in international collaboration.

5.2. Similarities in research topics

Here we used the terminology extraction and terminology variation identification components in TermWatch. This analysis is carried at two levels: topics (cluster labels) and topics contents (cluster contents).

By comparing cluster labels, we found topics that were common to authors from different geographic areas.

Table 1. Words specific in different data sets

<table>
<thead>
<tr>
<th>US_only</th>
<th>non_US</th>
<th>International</th>
</tr>
</thead>
<tbody>
<tr>
<td>catalog; class; classes; correlate;</td>
<td>circle; dot; early; equation;</td>
<td>absolute; based; discovery; discuss;</td>
</tr>
<tr>
<td>degrees; demonstrate; disk; due;</td>
<td>field; groups; method;</td>
<td>emission; estimate; galactic; imaging;</td>
</tr>
<tr>
<td>dwarfs; gamma; halo; halos;</td>
<td>models; obtained; order;</td>
<td>inflation; law; low; luminosity;</td>
</tr>
<tr>
<td>kpc; matter; previously; primary;</td>
<td>parameter; parameters;</td>
<td>magnitude; observations; photometric;</td>
</tr>
<tr>
<td>properties; quasar; quasars; relative;</td>
<td>power; ratio; ray; redshifts;</td>
<td>photometry; report; selected; simple;</td>
</tr>
<tr>
<td>satellite; satellites; sources; suggest;</td>
<td>scale; state; stellar; universe;</td>
<td>system; telescope; variation; wmap;</td>
</tr>
<tr>
<td>surveys; variables; wide;</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>US &amp; non_US</th>
<th>US &amp; International</th>
<th>non_US &amp; International</th>
</tr>
</thead>
<tbody>
<tr>
<td>background; cosmic; cosmological;</td>
<td>color; consistent; dwarf;</td>
<td>density; distribution; model; optical;</td>
</tr>
<tr>
<td>dark; microwave; observed; release;</td>
<td>function; high; mass; objects;</td>
<td>range; spectra</td>
</tr>
<tr>
<td>results; spectrum;</td>
<td>stars; type</td>
<td></td>
</tr>
<tr>
<td>Inter, USA, non_US</td>
<td></td>
<td></td>
</tr>
<tr>
<td>large; line; observed; present; redshift; show; star</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Overlap in cluster labels by geographic and cultural zones

<table>
<thead>
<tr>
<th>Nb_clusters</th>
<th>Non_US</th>
<th>US_only</th>
<th>Inter</th>
</tr>
</thead>
<tbody>
<tr>
<td>US, NonUS, Inter</td>
<td>552</td>
<td>6 (1%)</td>
<td></td>
</tr>
<tr>
<td>US vs Non_US</td>
<td>282</td>
<td>10 (4%)</td>
<td></td>
</tr>
<tr>
<td>US vs Inter</td>
<td>359</td>
<td>22 (6%)</td>
<td></td>
</tr>
<tr>
<td>Non_US vs Inter</td>
<td>403</td>
<td>29 (7%)</td>
<td></td>
</tr>
</tbody>
</table>
The overlap in cluster labels among the three data sets and between pairs of data sets is very low, thus pointing to significant differences in SDSS research across different geographic and cultural regions. It appears from the above figures that both US_only and non_US share more common points with international collaboration research than with each other as indicated by the very close overlaps (6% and 7%) with topics in international collaboration. The six labels common to all three geographic zones are: star formation rate, emission line, surface brightness, black hole, rest frame, large scale structure.

Comparison of the clusters contents obtained for each data set gives a measure of their overlap across the three data sets. Table 3 gives details of this comparison.

The proportion of overlap in cluster contents echoes the ones found among cluster labels. Thus, similarities are consistent whether we look at the topic labels alone or into their contents.

6. Conclusion
The results obtained here are encouraging for identifying the impact of the SDSS survey on the global research community of astronomers and the uniqueness of each cultural or geographic region in contributing to knowledge discovery and dissemination in this field. These results have shown that three geographical zones have distinct research preoccupations characterizing them but that each region (US and non-US) are brought together by international collaboration on some common research topics. This is remarkable considering that the terms were extracted automatically from the text fields of the bibliographic records. The systems have been able to automatically isolate the core set of shared knowledge among SDSS researchers worldwide without resorting to a human perusal of the publications which would be too time consuming.

Acknowledgement
The work reported here is in part supported by the National Science Foundation under Grant No. 0612129.

References
Social Tagging and Communities of Practice
Two Case Studies

Abstract
In investigating the use of social tagging for knowledge organization and sharing, this paper reports on two case studies. Each study examines how two disparate communities of practices utilize social tagging to disseminate information to other community members in the online environment. Through the use of these tags, community members may retrieve and view relevant Web sites and online videos. The first study looks at tagging within the Code4Lib community of practice. The second study examines the use of tagging on video sharing sites used by a community of French teenagers. Uses of social tagging to share information within these communities are analyzed and discussed, and recommendations for future study are provided.

Introduction
This paper investigates the use of social tagging in two disparate communities of practice. Social tagging is the act of applying tags in a social setting. Tags, as defined by Guy and Tonkin are “any word that defines a relationship between the online resource and a concept in the user’s mind” (2006). Social tagging, therefore, can be described as “the collective assignment of keywords to resources” (Trant 2006). The studies reported on in this paper look at how much the people tagging take into context the other people in their social space as defined by communities of context. The first study looks at the social tagging practices in a community of practice consisting of people who develop and implement computer software for and in libraries. The second study investigates the use of tagging on video-sharing sites by French high school students that participate in dance battles in a movement known as Tecktonik Killer. These two studies of disparate groups hope to add to the still developing body of Library and Information Science (LIS) literature on social tagging.

Etienne Wenger, who coined the term along with Jean Lave, defines communities of practice on his Web site as “groups of people who share a concern or a passion for something they do and learn how to do it better as they interact regularly” (Wenger n.d.). In order to be considered a community of practice, three characteristics are crucial. The first characteristic is that there must be an identifiable domain of interest. For example, Linux would be the domain of a group of people the meet monthly to discuss and learn about the Linux computer operating system. The second characteristic is a need for community. This could be a physical or virtual community but it is important that the community of practice engage in discussions and joint activities in order to help each other and share information. Third, members of a community of practice are practitioners and not merely observers. A group of people who regularly attend the opera together would not be considered a community of practice even though they have a shared interest.

Social Tagging in the Code4Lib community
The first case study explores the social tagging practices of the Code4Lib community of practice. The Code4Lib community consists mostly of systems librarians and software developers who work for libraries. Community members use various online methods to
share information and to learn from each other including Internet Relay Chat (IRC), the Planet Code4Lib blog aggregator, and an online journal.

Online communities of practice may use a specific tag on social bookmarking sites that enable Web-based resources to be shared, or “advertised,” within that community. By studying these tags, it is possible to investigate what influence, if any, the community has on the individual’s choice of tags, and to what extent the members consider community while tagging. If members of a community of practice do not tag differently for the community than those outside of the community, they may be using a social tagging tool, but any community or social benefit may simply be a by-product of personal tagging. Code4Lib is one such community that shares information by using a social tagging tool. Items bookmarked in del.icio.us, a popular social tagging Web site, with the tag ‘code4lib’ are shared in three ways with the community: 1) on a Web page created through the del.icio.us site, 2) on the Planet Code4Lib blog aggregator, and 3) on the Code4Lib IRC channel.

This study builds on a previous study that investigated tags made by fifteen Code4Lib members who bookmarked at least five items with the tag ‘code4lib’ on del.icio.us. Bookmarks tagged with ‘code4lib’ (community) were compared with items bookmarked by these members that did not include the ‘code4lib’ tag (personal) in order to see if different types of tags were used for the community. The tags were analyzed according to Golder and Huberman’s (2006, 203) seven mutually exclusive semantic categories: 1. Identifying what (or who) the item being tagged is about; 2. Identifying what the item being tagged is; 3. Identifying who owns the tag (ownership); 4. Refining categories; 5. Identifying qualities or characteristics of the item being tagged; 6. Self reference; and 7. Task organizing. Sets of tags were separated by individual user and were placed into categories based on the inclusion or exclusion of the ‘code4lib’ tag within the set. The total number of tags and the numbers of tags in each of Golder and Huberman’s categories were investigated to determine if there was any difference in kinds or number of tags used. While casual observation showed differences in how a few of the individuals tagged for themselves versus for the community, overall there was no significant difference in types of tags used in each set. The one statistically significant difference ($p < .01$) found was the number of tags applied in each set. The average number of tags used when the ‘code4lib’ tag was included was 3.70 compared to only 2.97 when the tag was not included. However, it was notable that when the tag ‘code4lib’ is excluded from the count of tags for these resources, the difference did not turn out to be statistically significant. This may mean the only difference was the inclusion of the community tag. (Tonkin et al. 2008)

**Method**

The tag ‘code4lib’ was searched for in the social bookmarking site del.icio.us in order to identify Web sites that someone, presumably community members, tagged with the term. Each Web site tagged was reviewed to see if it was about the Code4Lib community in some way (e.g. the Code4Lib conference Web site). If it was determined to be about the community, the site was excluded from the study. If it was not about Code4Lib, it was assumed that the site was tagged with ‘code4lib’ in order to share the site with other members of the community. Up to ten sets of user tags assigned to the Web sites that where identified this way were investigated. The most recent five sets of tags containing the term ‘code4lib’ were investigated. When there were less than five sets of tags containing ‘code4lib’ all of the tag sets containing the tag were examined. An equal number of tag
sets for the same resources from users that did not assign the tag ‘code4lib’ were also examined. For example, if there were four users who tagged a resource with ‘code4lib’ then four sets of tags from users that did not assign that tag were also explored. All sets of tags were manually investigated to see if any trends or differences could be observed. All totalled, one hundred sets of tags assigned to twenty-three separate resources were reviewed.

Results

The group of tag sets containing the ‘code4lib’ tag included a total of 288 tags (5.76 tags per set) while the group without the tag had only 221 total tags (4.41 tags per set). The difference between the numbers of tags by community taggers and non-community taggers was statistically significant when compared using a t-test ($p < .05$). Because the cases studied are not assumed to have a normal distribution, the Wilcoxon non-parametric test was also performed and yielded identical results. Despite a higher number of total tags in the Code4Lib group, the number of different tags was almost equal. The Code4Lib group assigned 123 unique tags while the non-community groups assigned 121 unique tags. Between the two groups, 189 unique tags where used. The tags were investigated to see if any difference in the types of tags used could be found by assigned each tag to one of Golder and Huberman’s seven kinds of tags. However, this did not produce any interesting results.

When the community identifier tag is excluded, the Code4Lib group still assigned, on average only 0.3 additional tags per item. T-tests and a Wilcoxon non-parametric test were performed on the data excluding the ‘code4lib’ community tag from the results. Without the community tag, no statistical significance was found. This result is in accord with the results found in the earlier study of tagging in the Code4Lib community of practice that found the only statistical difference may be explained by the inclusion of the community tag.

Discussion

One observation was that the Code4Lib community group was less likely to use a tag that defined something as ‘opensource’ (or related tags such as ‘open_source’ or ‘open-source’). ‘Opensource’ was one of four tags that were repeated the most times in the non-community group (eight times). The tag, or one of its variants, was only used three times by the Code4Lib community group. Considering the larger number of tags assigned by the Code4lib group this is an interesting observation. One possible explanation for this is that the Code4Lib community of practice works mostly with Open Source Software and thus, the observation of something being an Open Source project is not as worthy of a tag delineating that fact. Future studies may want to consider looking into the degree of specificity of tags within a community of practice.

The twenty-three Web sites reviewed all related to computer programming or libraries, and most related to both. One possible explanation for only a slightly larger number of tags assigned by the Code4Lib community group members do not feel any more ownership of these types of resources than the other taggers. Since they do not have ownership, they may not have an incentive to supply additional tags. Ownership, as we will see in the study of the Tecktonik killer community below, appears to imply a greater level of commitment to a resource, even in the online world. Thus, one of the challenges for organizations or communities of practice that hope to take advantage of social tagging will be to
somehow install a sense of ownership or some other incentive to get people to tag items more thoroughly. It appears that simply being a community of practice member does not necessarily imply a sense of ownership. As Lorcan Dempsey (2008) points out in his Weblog, one cannot simply add social networking tools to a Web site “and expect it to work well.” For example, think of all those empty Web-based forums that have been installed on sites over the years. Web sites with a larger user base such as del.icio.us may be more immune to the issue of ownership and incentive because of sheer scale, but it is unlikely that small communities can reach critical mass to develop interesting trends or folksonomies without cultivating a sense of ownership.

The Tecktonik killer community

Some counter-culture French high school students engage in dance battles in a movement called “Tecktonic killer” (Lazimi 2007). These teenagers dance to techno music and have a shared identity among members of the community. Along with competitions in the offline world, community members make videos of themselves dancing and post them online in video-sharing sites. Other community members can then comment on the technique in the videos, “favourite” the videos, and bookmark them using social bookmarking utilities like delicious. The competitive environment can remain, and community members can learn moves, vie for attention, and share thoughts with other community members. As a way of studying the extent to which the community of practice surrounding the Tecktonik killer dance phenomenon in France is using social tagging, tags supplied by both creators and users (those who posted videos and those who bookmarked them in del.icio.us) were investigated. Comments made by both groups were also collected, in part to understand better participation by French-speaking taggers, and in part to compare the degree of investment on the part of the two groups.

Research previous to the current study was carried out in October in 2007 right after the movement had started to be reported in French news sources. When the prior case study was performed, twelve videos posted to three video sharing sites used by the French (YouTube, Daily Motion, and video.fr) were investigated. Four types of videos were selected: exemplary, average, beginner, and parody. The tags for each type of video were counted and compared. The results of the study implied that French teenagers were familiar with social networking technologies and were participating in tagging activities in a way that was similar to other taggers. The French community in question has adapted to Web 2.0 in ways similar to other online communities. (Tonkin et al. 2008)

Author indexing and user indexing are traditionally perceived as being different approaches to the indexing task. The current research seeks to understand if taggers of Internet resources tag more when they perform the online equivalent of author indexing. “Endo-tags” or tags of one’s own material are assumed to form more robust tag sets due to the advertising function they incorporate. When supplying tags for their own content, endo-taggers are advertising their work to other members of the community. More tags make non-text resource more findable. In this study, we expect to see a relatively large number of endo-tags in video-sharing sites such as YouTube and Daily Motion.

“Exo-tags” or tags created by users of posted content will form less robust sets in comparison. Exo-tags are sufficient if they guide users back to a previously-found resource. It is assumed, for this research, that in social bookmarking sites like del.icio.us, exo-taggers are tagging the content of others for their own personal use. Exo-tags do not carry the
same advertising function as endo-tags, but rather serve as personal signposts in the online environment.

By using bookmarks, exo-tags, exo-comments, endo-tags and endo-comments for freely available video content, the current study seeks to understand more about the use of tags types in these online environments.

**Method**

Two alternate spellings of the dance phenomenon (tecktonic, tecktonik) were searched in the del.icio.us social bookmarking site as a means of identifying relevant resources. There is no well-known French-created social bookmarking site, but prior experience shows del.icio.us to be a site likely to be used for bookmarking French content. Search results were examined manually to ascertain the nature of the resource being bookmarked. To maintain continuity with the prior study, videos were the only bookmarked resources retained. The activity for a total of fifty-three del.icio.us users was collected. In some instances, multiple users had bookmarked a single URL in del.icio.us. All of the exo-tags, exo-comments, and data about date entered were harvested from del.icio.us and entered in a spreadsheet for computation.

On the video-sharing site for the corresponding video resources, the creator-supplied endo-tags and endo-comments, the number of times the resource was viewed, and the date it was posted (for purposes of disambiguation) were collected. Fifteen videos posted to video sharing sites were examined for this study.

**Results**

Endo-taggers supplied an average of 10.8 tags per video. Bookmarkers who supplied exo-tags gave an average of three per video. Eight bookmarkers did not supply tags; they bookmarked a URL for a Tecktonik killer video without supplying any exo-tags at all. The difference between average numbers of supplied endo- and exo-tags is extremely statistically significant when compared using a t-test ($p < .001$). Because the cases studied are not assumed to have a normal distribution, the Wilcoxon non-parametric test was also performed and yielded identical results.

**Discussion**

In his book *The Long Tail*, Chris Anderson (2006) describes the reputation economy that has developed online. Although they do not receive money for their efforts, generators of content who contribute to the online body of knowledge are trying to improve their own reputation. By tagging heavily, endo-taggers are seeking status while contributing to the online community of practice, and doing their best to draw attention to their contribution.

Conversely, exo-taggers in del.icio.us supply less tags. In part, this may be the result of being less invested in the creation of the content. Exo-taggers may also perceive that they are tagging for themselves instead of tagging for the community.

The issue of language also presents relevant cultural and community-based questions. Many of the del.icio.us bookmarks (n=15) supplied comments that were not in French. Some seemed to be written by native speaker of English who were discovering the originally French movement: “what happens when you mix Breakdance & Techno? This!” Other comments were less idiomatic and may not have been written by native speakers of English. Five del.icio.us exo-taggers supplied comments in French. One of these exo-
Taggers also supplied the English-language tag “music” to the set. The use of English, the lingua franca of the Internet, in the bookmarks implies an understanding of the international aspect of the movement. This is less obvious in the video-sharing sites where all but one of the comments was in French.

The importance of the community and the difference between core members and casual observers in this study should also be acknowledged. Endo-taggers are likely to be core members of the community who speak French, participate in offline dance battles, and compete in the online arena as well. The Paris TechnoParade of September 15, 2007 popularized the Tecktonik Killer style of dance and introduced it to international visitors and casual outside observers in France. We note that exo-taggers in this study largely tagged their resource in September 2007 or after, signalling a late-comer status in terms of the movement. One del.icio.us exo-comment in French calls the dance “ridicule” and another English comment describes it as “French people dancing in interesting ways to electronic.” These are not the comments one would expect from core members of the community who feel invested. Furthermore, exo-taggers are not necessarily invested in the advertisement of the movement and this may in part explain their reluctance to tag robustly.

Conclusion
While investigating the community approaches to tagging in case studies is useful, it does not provide an overview of the phenomenon. To see trends more clearly, it is advisable to look at larger data sets. However it should be remembered that with small or emerging communities, it is not always possible to get a larger data set to investigate and sometimes it is necessary to use smaller data sets in an attempt to understand what is happening at that time inside of the community.

These two different studies on disparate communities of practice show similarities in the way social tagging can be used for knowledge organization and information retrieval. The Code4Lib and Tecktonik Killer communities both take advantage of tagging techniques in order to share information with other community members. They both use tags not only to describe a document, but also to assign ownership and promote sharing. The Code4Lib community does this using Golder and Huberman’s (2006) ownership tags while Tecktonik Killer participants tag items with generic terms and proper nouns that advertise affiliations and community. Further study on larger communities would help determine how people consider their community of practice when tagging. Lastly, community taggers could themselves be interviewed and studied to understand better their motivations and their strategies of use.

References


Searching with Tags
Do Tags Help Users Find Things?

Abstract
This study examines the question of whether tags can be useful in the process of information retrieval. Participants were asked to search a social bookmarking tool specialising in academic articles (CiteULike) and an online journal database (Pubmed) in order to determine if users found tags were useful in their search process. The actions of each participants were captured using screen capture software and they were asked to describe their search process. The preliminary study showed that users did indeed make use of tags in their search process, as a guide to searching and as hyperlinks to potentially useful articles. However, users also made use of controlled vocabularies in the journal database.

Background
Classification is practiced by all humans with varying purposes and agendas. (Bowker and Star 1999) In traditional library classifications, the classifier was the cataloguer or indexer, an individual trained in the rules of information organisation to assign important information about the physical media and the subject matter of the content. On the web, the classifier has typically been the creator of the item, or an automated system collecting basic word frequency information to determine approximate topics. There has been a growing move to classify materials manually using consensus classifications created on the web by large groups of users tagging material on social bookmarking sites. Users are encouraged to add descriptive terms or tags to each bookmark. Tagging is the process of assigning a label to an item.

While other groups have been involved in creating index terms (for example, keywords with submitted articles), these keywords generally have a small circulation and are not widely used. Small scale indexing is common but generally covers a narrow range of topics and is specific to the article. Collaborative tagging systems such as CiteULike (http://www.citeulike.org) or Connotea (http://www.connotea.org) allow users to participate in the classification of journal articles by encouraging them to assign useful labels to the articles they bookmark.

Related Studies
Previous research in classification suggests that there is a distinct difference between user created or naive classification systems and those created by professional indexers. Beghtol (2003) While both systems employ subject based terms, users tend to employ terms that remind them of current or past projects and tasks, terms which could have little meaning to those outside their circle of friends, but are very meaningful to the user. (Malone 1984; Kwasnik 1999; Jones, Phuwanartnurak, Gill and Bruce 2005)

Mathes proposes that librarians embrace user assigned tags as a third alternative to traditional library classifications and author assigned keywords. (Mathes 2004) He and others also suggest that user classification systems would allow librarians to see what vocabulary users actually use to describe concepts and that this could then be incorporated into the system as entry vocabulary to the standard thesaurus subject headings. (Mathes 2004; Hammond, Hannay, Lund and Scott 2005)
Studies comparing the terminology used in tagging journal articles to indexer assigned controlled vocabulary terms suggests that many tags are subject related and could work well as index terms or entry vocabulary (Kipp 2006; Kipp and Campbell 2006; Hammond, Hannay, Lund and Scott 2005); however, the world of folksonomies includes relationships that would never appear in a traditional classification including time and task related tags, affective tags and the user name of the tagger. (Kipp 2007) These short term and highly specific tags suggest important differences between user classification systems and author or intermediary classification systems which must be considered.

Users searching online catalogues and databases often express admiration for the idea of controlled vocabularies and knowledge organisation systems, but find the process of searching frustrating. (Fast and Campbell 2004) Controlled vocabulary indexing has proven costly and has not proven to be truly scalable on the web. Can the user created categories and classification schemes of tagging be used to enhance retrieval in these new environments? Much speculation has been advanced on the subject but so far few studies have been done. A few projects are currently examining the combined benefits of professional and naive classifications. (Trant 2006; Allen and Winkler 2007; Quintarelli, Resmini and Rosati 2006)

The following study explores the usefulness of tagging for enabling retrieval by performing an information retrieval study on a social bookmarking system and a more traditionally classified database to study the usefulness of tags in the support of information retrieval. All information retrieval studies using controlled vocabulary searches contain an implicit evaluation of the effectiveness of classification terms. In such an evaluation it is important to evaluate not only the retrieval effectiveness of the search term, but also how long it took the user to think of using this term in this context and whether or not the user thought the term was useful and accurate.

This study aims to explore those questions in a new context. Proponents of tagging and social bookmarking often suggest that tags could provide at worst an adjunct to traditional classification systems and at best a complete replacement for such systems. (Shirky 2005) A method for testing the usefulness of a classification system for enabling retrieval is to perform an information retrieval study on a social bookmarking system to study the usefulness of tags in the support of information retrieval.

One way to examine the potential uses of tags in the search process is to compare the search experience between social bookmarking tools and other methods of information retrieval such as retrieval via controlled vocabulary or retrieval via free text search.

Research Questions

- Do tags appear to enhance retrieval? Do users feel that they have found what they are looking for?
- How do users find searching social bookmarking sites compared to searching more classically organised sites? Do users think that tags assigned by other users are more intuitive?
- Do tagging structures facilitate information retrieval? How does this compare to traditional structures of supporting information retrieval?

Methodology

A preliminary study was conducted using volunteer searchers. Participants are currently being recruited to continue the study. The searchers were asked to search Pubmed (an
You are a reference librarian in a science library. A patron approaches the reference desk and asks for information about the application of knowledge management or information organisation techniques in the realm of health information. The patron is looking for 5 articles discussing health information management and is especially interested in case studies, but will accept more theoretical articles as well.

Screen capture software, a think aloud protocol and an exit interview were used to capture the impressions of the users when faced with traditional classification or user tags and their usefulness in the search process. While information concerning the usability of the systems themselves for searching may be of interest, data collection will be focused on a comparison of the terms entered by the participants.

<table>
<thead>
<tr>
<th>Activity Description</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Welcome</td>
<td>2–3 minutes</td>
</tr>
<tr>
<td>Introduction to session</td>
<td>5–7 minutes</td>
</tr>
<tr>
<td>First search task (CiteULike or Pubmed)</td>
<td>15 minutes</td>
</tr>
<tr>
<td>Post search discussion</td>
<td>15 minutes</td>
</tr>
<tr>
<td>Conclusion</td>
<td>3–5 minutes</td>
</tr>
</tbody>
</table>

Each participant searched for information using both the traditional on-line database with assigned descriptors and a social bookmarking site. Participants were asked to perform the searches in the order specified so that their use of a social bookmarking site first versus an on-line database could be randomised.

Participants selected their own keywords for searches on both tools. Participants were asked to search until they had located 5 articles that appeared to match the query and assign relevance score to article based on an examination of available metadata. At the end of each search, participants were asked to make a list of what terms they would now use if asked to search for this information again. Participants did not have access to their initial set of search terms at this time to eliminate the learning effect.

Three sets of data were thus available for analysis: sets of initial and final keywords selected by the user, the recording of the search session and think aloud, and recorded exit interviews after the search session. Each set of data can be analysed to examine user impressions of the search process from the perspective of the keywords (tags or index terms respectively). Additionally, keywords and tags chosen by users will be compared and examined to see how or whether they are related.
Results
Preliminary results from the study show that users tended to prefer the search experience on the system used first, regardless of previous experience with either system. Further interviews are required to determine if this trend continues.

All users used multi word keywords initially, which is unsurprising as they are in training to be librarians. At the end of the search process, when users were asked to generate a new list of keywords they would now use for the search, a majority of the users separated their list of final keywords by tool, despite the fact that they were asked for only one list.

<table>
<thead>
<tr>
<th>Table 2. Initial Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Keyword</strong></td>
</tr>
<tr>
<td>knowledge management</td>
</tr>
<tr>
<td>information organisation</td>
</tr>
<tr>
<td>case studies/&quot;case stud&quot;</td>
</tr>
<tr>
<td>health information</td>
</tr>
<tr>
<td>health information management</td>
</tr>
<tr>
<td>consumer health</td>
</tr>
<tr>
<td>consumer health management</td>
</tr>
</tbody>
</table>

The four most commonly chosen terms were: knowledge management, information organisation, case studies/"case stud" and health information. Each of these terms is directly from the initial text of the information need.

<table>
<thead>
<tr>
<th>Table 3. Final Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Keyword</strong></td>
</tr>
<tr>
<td>CiteULike</td>
</tr>
<tr>
<td>knowledge management/km</td>
</tr>
<tr>
<td>information management</td>
</tr>
<tr>
<td>health-information/health information</td>
</tr>
<tr>
<td>case studies/&quot;case stud&quot;</td>
</tr>
<tr>
<td>health-informatics</td>
</tr>
</tbody>
</table>

The most commonly used keyword, by far, was knowledge management. This term comes directly from the information need (described above) and is in keeping with previous information retrieval studies where users tended to select terms from the text for search. (Oppenheim, Morris and Mc Knight 2000) Information management (also commonly used), could be seen as a modification of knowledge management to fit the terminology of a different group of users who prefer the term information management. Another commonly chosen term was health information (from the information need). Both information management and health information were 2nd and 3rd most popular for CiteULike and Pubmed (or vice versa). While users considered their initial keyword sets to have been incomplete, they tended to choose the same or very similar terms as their suggestions for good search terms to use in order to produce better results. This suggests that their initial search terms were well chosen and matched closely those chosen by users tagging articles in CiteULike, but also came close enough to terms used in the Medical Subject headings used in Pubmed (or its entry vocabulary) or terms used by authors whose works are pub-
lished in Pubmed for good results to be retrieved. Users selected from 3–5 keywords for both lists, but one user found the Pubmed descriptor “information management” to be the best keyword possible and suggested that users would be better to use this MeSH heading and browse the results.

Conclusions
The preliminary study showed that users did use the tags to aid in the search process, selecting tags to see what articles would be returned. They also used the tags as a guide to suggest further search terms, suggesting that users do indeed pay attention to subject headings and metadata if it fits a pattern they recognise or makes sense in the context of their existing knowledge on the subject.

Users generally used the same number of keywords for both lists, though most insisted on dividing the final keyword list up by tool. Despite this, the 3 most commonly used terms were the same in each case and knowledge management was generally selected as a useful term for each tool. Interestingly, users tended to say they preferred searching the tool they used for the first search, regardless of prior experience. Further study is required to see if this phenomenon holds true.

It is expected that the continuing study will provide additional insight into user’s choices of preliminary keywords for searching as well as participant insights into the process of searching via the tags or controlled vocabulary.

References


Creating Pathways to Memory
Enhancing Life Histories through Category Clusters

Abstract
For individuals whose memory and language are intact, making sense of unfamiliar information or objects is a process of matching what is unknown, to what is known through previous learning or experience. The unfamiliar is linked to clusters or categories of the familiar, identifying what is “like” or “nearly like” and excluding all others (De Mey 1982). Most commonly, these are categories on which there is general agreement, sometimes collocated under established terms, labels, or shared naming devices. Classification systems are built on the basis of shared understandings of human knowledge and culture. When memory and/or language are impaired, how does such contextualizing and categorizing occur? Since perception is individual, can a person with cognitive impairment “make sense” of information, an object, a situation, using alternative modes of expression that are less or not language-dependent? This paper reports on preliminary results from a pilot study undertaken as part of exploratory mixed methods research examining the sense-making, sorting, categorization, and recall strategies of individuals with mild cognitive impairment (MCI) resulting from early stage dementia.

Introduction and Background to the Study
Alzheimer’s Disease International (ADI) describes dementia as a condition that affects memory, thinking, behaviour, and emotion. Declining memory, especially short-term memory, is the most common early symptom of dementia. Other symptoms include difficulty performing familiar tasks, disorientation to time and place, poor or decreased judgment, and changes in personality. Alzheimer’s disease is the most common cause of dementia (ADI 2007). As part of the Cognitive and Emotional Health Project (CEHP) (USA), a critical evaluation study committee was charged with assessing the state of epidemiological research on demographic, social, and biological determinants of cognitive and emotional health (Hendrie et al. 2006). Their review of large, longitudinal cohort studies (n=36) noted that the majority were disease-focused, with research on healthy brain aging lagging noticeably behind. Descriptions of drug and (other) medical interventions to maintain cognitive health or prevent decline were well represented in the literature. Possible lifestyle interventions received less attention, but, nonetheless, revealed several protective factors, including higher education levels, occupational attainment, higher socioeconomic status, specific cognitive activities (e.g., playing board games, singing, playing a musical instrument, reading), social engagement, emotional support, instrumental mastery and self-efficacy, resilience, and vitality (Bain 2006; Hendrie et al. 2006; Jedrziewski, Lee, and Trojanowski 2005; Morrison-Bogorad, Cahan, and Wagster 2007; Cherry and Reed 2007). The critical review committee concluded that, “There is now widespread public interest in developing strategies to maintain or enhance cognitive and emotional health in the elderly” (Hendrie et al. 2006: 26). Emphasizing the need for future research that considers brain health maintenance, as well as disease prevention, the committee recommended that biomedical investigators join forces with other disciplines, such as social sciences and bioethics, to “change the paradigm of successful cognitive and emotional aging.” (Hendrie et al., 2006: 28).

In a documentary on memory and the brain aired on the Canadian Broadcasting Corporation, or CBC Radio program, Ideas, Marilyn Powell observed that, “Memory is lost because we cannot find our way back to it.” (Powell, Oct. 3, 2007).
As Figure 1 illustrates, as the human brain receives sensory input, that information is stored temporarily in the hippocampus, and slowly moved, or ‘consolidated’, into the neocortex. Memory consolidation is the process by which recent or short-term memories, comprised of sensory fragments, are crystallized into long-term memory. The more numerous or frequent the recall, or ‘rehearsal’ of that information (a process one might associate with ‘memorizing’), the more that memory is reinforced, or ‘reconsolidated’. With the stimulus of a cue or trigger, stored fragments from our experiences of people, places, events, and objects are reconstructed as part of our personal reminiscences or life histories. Mild, moderate or severe cognitive impairment, whether from disease, trauma, disability, or genetic predisposition, can disrupt the reassembly process, impeding access to memory.

Tulving (1972) was first to identify two primary types of memory, namely episodic, and semantic. Episodic memory involves recalling personal episodes (“autobiographical memory”), events, or experiences that are specific to time and place. In contrast, semantic memory, loosely described as the sum total of everything we know, “… includes knowledge of facts, concepts, moral and social cognition, as well as words and their meaning.” (Kazui et al. 2003: 983). It is shared culturally, is not time-dependent, and supports more complex “sorting” activities of categorization and association. While it is possible to have semantic memory without episodic memory, the reverse is not true — underscoring the importance of context or perceived meaning to the experience of day-to-day people, places, events, and things (Powell 2007). As personal semantic memories are repeatedly recalled in different contexts, their link to specific episodic information obtained in earlier contexts weakens while their semantic richness increases.

For those experiencing mild cognitive impairment (MCI) through the onset of early stage dementia, brain injury, or other neurodegenerative causes, drug and alternative clinical interventions have been devised to slow decline in cognition, or, as in dementia, specific memory loss. The theoretical and applied research literatures of gerontology, nursing, social work, and cognitive psychology/cognitive science describe approaches to reinforcing an individual’s “cognitive reserve,” defined as, “increased numbers of synapses, or and increased ability of the brain to cope with physiological insults, e.g., as a result of neural plasticity or neural compensation” (Bain 2006: 247). In short, healthy aging nerve cells can regenerate (Khachaturian 2007). Intelligence, educational level, occupation attainment, and (cognitive) leisure activities have been associated with reduced risk of dementia and cognitive decline, and are, consequently, used as proxy measures of cognitive reserve (Bain 2006). Specific clinical interventions may include music
programming (Topo et al. 2004), art therapy (Motram 2003), life story/life review work (Haight, Gibson, Michel, 2000; McKeown, Clarke, and Repper, 2005; Meininger 2005; Parker 2001), and reminiscence therapy (Cohen and Taylor 1998; Kim et al., 2006; Moss et al., 2002). Websites, such as BiFolkal and Memory Lane, offer products (videos, audio CDs, memory books, photographs and slides, tactile objects) and programs to “support remembering and reminiscing in people with mild cognitive impairment and mild to mid-stage AD” (Dishman and Carrillo 2007). Multimedia DVDs that store and play client (patient) biographies are currently under development.

With its focus on linking individuals with a particular information need to appropriate, specific information objects knowledge organization as a discipline seems well-situated to partner with researchers and clinicians focused on “healthy brain” initiatives (Cherry and Reed 2007). While there is a body of Library and Information Science/Information Studies (LIS/IS) literature dealing with developing collections to support research in gerontology and life course (aging), as well as to address the information needs of older adults, there is scant reference to LIS/IS research focused on aspects of mild cognitive impairment, dementia, or Alzheimer’s. Bath and Bouchier (2003) have developed tools for evaluating Internet information resources specific to Alzheimer’s. Nonetheless, this overall paucity of research stands in sharp contrast to, for example, studies of children’s design of Internet portals to satisfy their information needs (Large et al., 2007; 2006), or to explorations of information use among less advantaged or marginalized populations (Chatman 1992; 1990; 1987). Studies concerning information needs, uses, and search tools among adult populations, generally, are too numerous to cite; this segment of the broad demographic is well served by LIS/IS research.

**Study Objectives and Research Questions**

Building on previous research on the use of category clusters for grouping information in topic map displays for use in language-neutral situations (Howarth and Miller 2006a; 2006b), a broader exploratory, mixed methods study has been undertaken to examine the sense-making, sorting, categorization, and recall strategies of individuals experiencing mild cognitive impairment (MCI) resulting from early stage dementia. Specifically, the overall study seeks to provide baseline data for the following three research questions: (1) How do individuals diagnosed with MCI make sense of, organize, and categorize their childhood and adult life histories? (2) Are there differences between whether a standard (generic or icon) representation, or a personalized representation of a category is used to gather together an individual’s memories in a life history? and (3) How and how well do multi-modal expressions of information, e.g., music, drawings, physical objects, movement, serve as alternate means of creating context, thereby facilitating collocation of like information (category clusters) to assist with finding, organizing, and interpreting life memories?

**Pilot Study: Methods, Interpretive Frameworks, Preliminary Findings**

This ISKO 2008 conference paper reports on preliminary outcomes from a pilot study undertaken to examine the “sense making,” sorting, categorizing, and recall strategies of four residents living in an assisted-care facility in Toronto, Canada, for individuals with early stage dementia. Using the interpretive biographical methodology (Bruce et al. 2002; Surr 2006), unstructured tape-recorded interviews were conducted to gather individual narratives covering childhood, family, home, and adulthood. Follow-up sessions were
held to confirm recollections, to put together life history books, to identify/derive meaningful categories and category names for groupings of experiences, and to determine with participants effective cues or clues for recalling memory (e.g., photograph as substitute; musical cue; etc.). At a subsequent session, each of the four residents were provided, in turn, with a standard (generic or icon) representation of a life history category (e.g., commercial memory-evoking products use a photograph of a wedding to stimulate recollections of an individual’s own marriage), then with a personally-customized category cue and asked to talk about an associated memory or memories. The modality of cues, though aligned, were varied — i.e., an audio recording of the “wedding march”; the participant’s self-determined audio cue — for example, a song that the participant connected with the specific event or time period in which the event occurred.

De Mey’s (1982, 4) observation that, “The central point of the cognitive view is that any such information processing, whether perceptual (such as perceiving an object) or symbolic (such as understanding a sentence) is mediated by a system of categories or concepts which for the information processor constitutes a representation or a model of his world.” [emphasis in original] offered a useful conceptual framework for considering “sense-making” in the context of mild cognitive impairment, and for addressing the first research question in this study. Comparing the use of standard or personalized category representations (Research question #2) was informed by the work of Joubert et al. (2004) who determined that memory recall was enhanced by visual material, such as family pictures, rather than by verbal descriptions. Finally, Davidoff and De Blesser’s (1994: 2), findings that “Naming of real objects presented either visually or for tactile inspection was reliably better than naming photographs or line drawings of the same objects” provided a framework for exploring the use of multi-modal cues as alternate representations for participant categories (Research question #3).

The identification/derivation of meaningful categories and category names for groupings of experiences proved to be highly individualized, and sometimes problematic for the pilot study participants. This may have been as a result of categorizing tending to the abstract, where recall more usually required the concrete. The process of aggregating, which necessitated decisions as to (1) what to include in, and exclude from a grouping, and (2) what to name or label the aggregate, was, at times, challenging, with difficulties experienced relative to ambiguous or overlapping experiences. For example, travel featured prominently in the childhood of one participant. In putting together her life history, this participant sometimes favoured the category, ‘places’, while other times choosing to associate a travel experience with ‘people’ (normally family). In recall, probing for memories by use of the category, ‘places’, evoked recollections of geographic locations where the participant had lived or vacationed (e.g., a family cottage), rather than travel destinations and experiences. In contrast, cueing (i.e., in this instance, by means of a map) specific locales, such as “Norway”, resulted in quite thick descriptions of a trip there. Likewise, recall by means of the aggregated category, ‘people’, did not lead to mention of particular travel experiences. While (1) initial grouping of like experiences, and (2) the identification/derivation and naming of categories were completed as activities in the course of creating life histories, recalling particular memories associated with a labelled aggregate was more problematic, sometimes disconnected. The use of named categories, both for grouping and recalling like experiences proved less reliable for reinforcing memory than did different modal representations identified and favoured by a participant.
Preliminary findings suggest further that, standard, iconic representations more usually resulted in a generalized memory (i.e., talking about weddings as events), or a personal memory derived indirectly from the cue (i.e., talking about weddings, in general, followed by the individual’s experience, in particular), as opposed to specific, detailed recollections of the event. This finding carried over to multi-modal clues that had been self-selected by participants. Personalized category cues thus provided more direct access to life histories, and, as compared with standard icons, little or no confusion as to the meaning of the representation, per se. Outcomes may suggest that, while generic representations may evoke semantic memory, self-determined category markers may be necessary to facilitating episodic memory.

Conclusion
The design of the larger research project necessitated a pilot study for validating approaches, and identifying directions in which preliminary findings from examination of the sense-making, sorting, categorizing, and recall strategies of four participants diagnosed with MCI might lead. While outcomes are preliminary and warrant further scrutiny, they do reinforce that, while day-to-day people, places, events, and objects can exist in common, perception or the experiencing of them is individual, and contextualized uniquely by each observer or participant. A growing body of literature confirms that the human brain can adapt to trauma caused by injury or disease, that it can regenerate, and compensate. Neural plasticity ensures that other parts of the brain can “learn” and assume some function(s) that affected areas can no longer do. Research has also suggested that reinforcing cognitive reserve can slow specific memory loss resulting from the onset of dementia. The pilot study offers preliminary insights into the role that sense-making, sorting, and categorizing play in the recall and reconstruction of long-term memories, and the degree to which alternate modes of evoking aggregated or single memories can be used to compensate for the loss of other reminiscence strategies. At the least, the research will yield baseline data in a largely unexplored area of knowledge organization. More important to those with MCI and their care-givers, if information targeted to the particular needs of an individual can be delivered in a meaningful way, in a manner that makes sense to its recipient, in whatever mode expressed and communicated, then prospects for continued independence as well as opportunities for extending memory and language capabilities may be optimized.

References


Large, Andrew et al. 2006. “‘Bonded Design’: a novel approach to intergenerational information technology design.” *Library and Information Science Research* 28(1): 64–82.


Machine Versus Human Clustering of Concepts Across Documents

Abstract
An automated method for clustering terms/concepts from a set of documents on the same topic was developed for the purpose of multidocument summarization. The clustering method makes use of a combination of lexical overlap between multiword terms, syntactic constraints and semantic consideration based on a manually constructed taxonomy to generate hierarchically organized clusters of terms. This study evaluates the machine-generated clusters by calculating the proportion of overlap with two sets of human-generated clusters for 15 topics. It was found that the overlap between machine-generated clusters and individual human-generated clusters are higher than that between two human-generated clusters. A qualitative analysis of the human clustering found that clusters formed are either semantic-conceptual based or lexical based (similar to machine clustering). The semantic-conceptual based clusters that were formed tended to be different for different human coders. This has raised questions about whether machine-generated clustering can be evaluated by comparing with human clustering.

Introduction
This paper reports a study of machine and human clustering of concepts across documents. The study is carried out in the context of multi-document summarization research—to develop an automatic method to summarize a set of related documents on a particular topic. Multi-document summarization involves identifying common information found in multiple documents, relations between the pieces of information, as well as unique information found in individual documents. Our approach to multi-document summarization focuses on extracting important terms/concepts in the documents and identifying the relations specified between the concepts. Thus an important step in the summarization method involves clustering similar or related concepts.

We adopt a method for clustering concepts using a “global taxonomy” developed manually based on a corpus, in combination with a “local taxonomy” or hierarchical structure of terms constructed from the set of related documents to be summarized. In this paper, we report the results of an evaluation that compares the clusters generated by the automatic method with clusters constructed by human coders. The objectives are to find out:

1. How “good” the machine-generated clusters are compared to human-generated clusters
2. The characteristics of human-generated clusters.

Clustering and categorization are a fundamental human behavior, and though there have been many studies of human categorization in the field of cognitive psychology, they have focused on categorization of common objects and concepts. We have not found any work by information science researchers on human clustering of terms and concepts taken from documents. The kind of human clustering research that is closest to ours is the card sorting studies sometimes carried out to develop menu hierarchies and taxonomies for organizing Web sites and information system interfaces. Bar-Ilan and Belous (2007) studied how children organized subject categories taken from Web directories using card sorting. Faiks and Hyland (2000) used card sorting to find out how users would organize the concepts for a digital library help system. However, most articles have focused on the practical aspects of card sorting methodology, analysis and use.
We view document summarization as involving the fundamental steps of identification, clustering and categorization of concepts in the documents. From this point of view, knowledge of how humans cluster concepts in documents will provide insights for the design of good automatic summarization methods.

This study makes use of dissertation abstracts (from the Dissertation Abstracts International database) in the field of sociology as the corpus.

**Automatic Clustering Method**

Most clustering algorithms are based on analyzing the positions of items in a multidimensional space and the distances between them. Items that are close together in the multidimensional space are assigned to the same cluster. The clustering method used in this study is not a typical distance-based clustering algorithm. It has two phases. In the first phase, similar terms are clustered together based on common words found in the terms and organizing the terms into a hierarchy with the shorter terms (representing broader concepts) higher in the hierarchy. Various syntactic constraints are also used to select the terms for the cluster. The second phase involves assigning term clusters to subject areas based on a manually constructed “global taxonomy.”

The clustering process has following steps:

1. **Segment multiword terms.** Multiword terms occurring in the documents are segmented into 1, 2, 3, 4 and 5-word terms. A set of predefined syntactic patterns is used to filter out valid terms for clustering.

2. **Construct term chains.** Starting from each 1-word term, a list of term chains are constructed by linking it level by level with other multi-word terms in which the single word is used as a head noun. Each chain is constructed iteratively by linking a shorter term with the longer terms containing the shorter term. The shorter terms represent the broader concepts at the higher level whereas the longer terms represent narrower concepts.

3. **Build cluster tree.** All the chains sharing the same root node (1-word term) are linked to form a hierarchical cluster tree (see illustration in Figure 1). Each cluster tree uses the root node as its cluster label.

4. **Assign the clusters to subject areas.** The cluster labels are mapped to subject areas using a taxonomy. The taxonomy was manually constructed based on an analysis of terms extracted from a sample of over 3000 abstracts in the Dissertation Abstracts International database indexed under sociology in 2001. Frequently occurring terms were manually assigned to subject areas taken from two thesauri—the UNESCO Thesaurus, and the Humanities and Social Science Electronic Thesaurus (HASSET).

More details of the clustering algorithm can be found in Ou, Khoo and Goh (2005 & 2008).
Figure 1. A cluster tree containing five term chains

Table 1. Topics used in the comparison of machine and human clustering

<table>
<thead>
<tr>
<th>Doc. Set ID</th>
<th>Topic</th>
<th>No. of retrieved docs</th>
<th>No. of selected docs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>attachment and marriage</td>
<td>483</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>racial socialization</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>adolescent suicide</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>demographic transition</td>
<td>16</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>school success</td>
<td>21</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>intermarriage</td>
<td>25</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>unemployment</td>
<td>129</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>health policy</td>
<td>36</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>family planning</td>
<td>45</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>mass media</td>
<td>49</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>rural development</td>
<td>64</td>
<td>5</td>
</tr>
<tr>
<td>12</td>
<td>juvenile delinquency</td>
<td>71</td>
<td>5</td>
</tr>
<tr>
<td>13</td>
<td>welfare reform</td>
<td>89</td>
<td>5</td>
</tr>
<tr>
<td>14</td>
<td>substance abuse</td>
<td>106</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>childhood sexual abuse</td>
<td>57</td>
<td>5</td>
</tr>
<tr>
<td>16</td>
<td>rural development</td>
<td>64</td>
<td>5 + 5 (from set 11)</td>
</tr>
<tr>
<td>17</td>
<td>juvenile delinquency</td>
<td>71</td>
<td>5 + 5 (from set 12)</td>
</tr>
<tr>
<td>18</td>
<td>welfare reform</td>
<td>89</td>
<td>5 + 5 (from set 13)</td>
</tr>
<tr>
<td>19</td>
<td>substance abuse</td>
<td>106</td>
<td>5 + 5 (from set 14)</td>
</tr>
<tr>
<td>20</td>
<td>childhood sexual abuse</td>
<td>57</td>
<td>5 + 5 (from set 15)</td>
</tr>
</tbody>
</table>
Table 2. Overall similarity between two coders and between the system and each coder

<table>
<thead>
<tr>
<th>Doc. set ID</th>
<th>Similarity between two coders</th>
<th>Similarity between system and coder 1</th>
<th>Similarity between system and coder 2</th>
<th>Avg. similarity between system and the coder</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.111</td>
<td>0.149</td>
<td>0.533</td>
<td>0.341</td>
</tr>
<tr>
<td>2</td>
<td>0.092</td>
<td>0.093</td>
<td>0.137</td>
<td>0.115</td>
</tr>
<tr>
<td>3</td>
<td>0.144</td>
<td>0.106</td>
<td>0.137</td>
<td>0.122</td>
</tr>
<tr>
<td>4</td>
<td>0.196</td>
<td>0.236</td>
<td>0.274</td>
<td>0.255</td>
</tr>
<tr>
<td>5</td>
<td>0.170</td>
<td>0.259</td>
<td>0.085</td>
<td>0.172</td>
</tr>
<tr>
<td>6</td>
<td>0.092</td>
<td>0.199</td>
<td>0.123</td>
<td>0.161</td>
</tr>
<tr>
<td>7</td>
<td>0.438</td>
<td>0.085</td>
<td>0.171</td>
<td>0.128</td>
</tr>
<tr>
<td>8</td>
<td>0.274</td>
<td>0.182</td>
<td>0.131</td>
<td>0.157</td>
</tr>
<tr>
<td>9</td>
<td>0.147</td>
<td>0.080</td>
<td>0.123</td>
<td>0.102</td>
</tr>
<tr>
<td>10</td>
<td>0.172</td>
<td>0.321</td>
<td>0.452</td>
<td>0.387</td>
</tr>
<tr>
<td>11</td>
<td>0.394</td>
<td>0.374</td>
<td>0.602</td>
<td>0.488</td>
</tr>
<tr>
<td>12</td>
<td>0.258</td>
<td>0.227</td>
<td>0.550</td>
<td>0.389</td>
</tr>
<tr>
<td>13</td>
<td>0.035</td>
<td>0.040</td>
<td>0.403</td>
<td>0.222</td>
</tr>
<tr>
<td>14</td>
<td>0.216</td>
<td>0.244</td>
<td>0.698</td>
<td>0.471</td>
</tr>
<tr>
<td>15</td>
<td>0.234</td>
<td>0.172</td>
<td>0.433</td>
<td>0.303</td>
</tr>
<tr>
<td>Average</td>
<td>0.198</td>
<td>0.254</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.136</td>
<td>0.154</td>
<td>0.205</td>
<td>0.180</td>
</tr>
<tr>
<td>17</td>
<td>0.258</td>
<td>0.236</td>
<td>0.588</td>
<td>0.412</td>
</tr>
<tr>
<td>18</td>
<td>0.163</td>
<td>0.205</td>
<td>0.274</td>
<td>0.240</td>
</tr>
<tr>
<td>19</td>
<td>0.173</td>
<td>0.235</td>
<td>0.300</td>
<td>0.268</td>
</tr>
<tr>
<td>20</td>
<td>0.038</td>
<td>0.081</td>
<td>0.083</td>
<td>0.082</td>
</tr>
<tr>
<td>Average</td>
<td>0.154</td>
<td></td>
<td></td>
<td>0.236</td>
</tr>
</tbody>
</table>

**Human Clustering of Concepts**

For the comparison of machine versus human clustering, 15 topics in the field of sociology were haphazardly selected from the titles of dissertations. The topics are listed in Table 1. For each topic, a set of PhD sociology dissertation abstracts were retrieved from the Dissertation Abstracts International database using the topic as the search query, and five abstracts were selected to form a document set. Moreover, for five of the topics (Document set ID 11 to 15), an additional five abstracts were selected for each topic and combined with the previously chosen five abstracts to form a second bigger document set. These bigger document sets were used to examine the difference in clustering between small (5-document) and bigger (10-document) sets. For each dissertation abstract, terms were extracted from the research objectives and research results sections automatically by the system.

The human coders were asked to read the abstracts and then examine the list of terms extracted by the system. They then sort similar concepts into clusters and assign a label to each cluster. The coders were not compelled to sort all the terms listed, but could leave out terms that they did not perceive as forming clusters. It was felt that this approach was more naturalistic. The human coders were social science graduate students at Nanyang Technological University, Singapore. Each document set was coded by two human coders, and each human coder coded three 5-document sets and one 10-document set on different topics.
Results

On average, the coders created 4 clusters for the 5-document sets and 7 clusters for the 10-document sets, whereas the system created 17 clusters for the 5-document sets and 32 clusters for the 10-document sets. Thus, the system created many more clusters than the human coders. The average size of the clusters created by the coders was 12 concepts for the 5-document sets and 15 concepts for the 10-document sets, whereas the average size created by the system was 6 concepts for the 5-document sets and 7 concepts for the 10-document sets. The average size of the system-created clusters was almost half that of human-created clusters. The system created much smaller clusters than the human coders.

For a 5-document set, on average, the system clustered 68 concepts in total whereas the coders clustered 46 concepts in total. For a 10-document set, on average, the system clustered 149 concepts in total whereas the coders clustered 85 concepts in total on average. This indicates that the system “worked harder” and clustered a bigger number of concepts than the human coders. When the size of the document set increased, the difference in the number of clusters between system and humans increased.

For each document set, two sets of clusters (codings) were created by two human coders and one set of clusters was created by the system. We calculate the similarity between two sets of clusters, set 1 and set 2, using the similarity measure employed by Macskassy et al. (1998). First, all possible pairs of terms in each cluster are identified, i.e. we obtain the set of same-cluster-pairs of terms for set 1 and for set 2. We then determine how many pairs are common between set 1 same-cluster-pairs and set 2 same-cluster-pairs, and the total number of unique pairs in the union of set 1 same-cluster-pairs and set 2 same-cluster-pairs. The overall similarity between set 1 and set 2 is calculated using the following formula:

\[
\text{Number of common same-cluster-pairs between set 1 and set 2} \div \text{Total number of unique pairs obtained from set 1 and set 2}
\]

The overall similarity between two coders and between the system and each coder for each of the 20 document sets is given in Table 2. The similarity values range from 0.04 to 0.44 across the 20 document sets. The average similarity obtained for the 20 document sets was a low 0.19. This means that clustering is a very subjective operation. The average inter-coder similarity obtained for the 10-document sets (0.15) was lower than that for the 5-document sets (0.20), indicating that human clustering becomes more difficult when the size of the document sets increases. However, there is no difference between the system-coder similarity for the 10-document sets (0.24) and that for the 5-document sets (0.25). This indicates that the system’s clustering does not become worse when the size of the document sets increases.

Among the 20 document sets, the average similarity between the system and the coder (0.25) was surprisingly higher than the average inter-coder similarity (0.19). A Paired-Samples T-test found that the system-coder similarity scores were significantly better than the inter-coder similarity scores \(p = .04\).

This result is counter-intuitive. A qualitative analysis of the human-generated clusters is being carried out to understand the characteristics of the human clusters and how they differ from the machine-generated clusters. Our preliminary observations are as follows. Both human and machine clustering can take two approaches:

1. A lexical approach to identifying similarities between two terms by locating words that are common in the two terms
2. A semantic-conceptual approach based on domain knowledge, and conceptual understanding of the common issues in the domain area.
The machine is best at using the lexical approach because it is tireless and thorough in carrying out word matching. Human coders were expected to adopt the semantic-conceptual approach. However, we found variations in the clustering approach used by different coders—some adopted mainly a semantic-conceptual approach, others used mainly a lexical approach, and yet others creating both types of clusters.

For example, one coder created the following semantic-conceptual clusters for topic 15 “child sex abuse”:
- Victims (groups of people)
- Characteristics of victims
- Family factors
- Mediating factors
- Clinical interventions
- Eating disorders (a consequence)
- Psychological disorders (consequence)
- Context of the abusive experience
- Personality tests/psychological measurements

Another coder created the following lexical-based clusters for topic 11 “rural development”:
- Rural development
- Participation
- People’s participation
- Farmer’s participation
- Community participation

It was observed that for most of the semantic-conceptual clusters, their cluster labels refer to one of the following types of concepts:
- Groups of people
- A type of effect or consequence
- A type of factor that produced the effect
- An issue that is important in the field
- A sub-topic of the main topic.

It is hypothesized that whether a coder uses the lexical or the semantic-conceptual approach of clustering depends on the conceptual ability of the coder, domain and research knowledge, and perhaps whether the coder is prepared to make the mental effort of conceptualizing the categories. If two coders use different clustering approaches, then of course their clusters are likely to be different.

Furthermore, many terms in the documents do share common words, and thus the lexical approach used by the automatic method manages to construct many clusters and appears to be quite effective in matching at least some of the human-generated clusters (especially the lexically-based clusters).

For five of the topics, two coders were asked to cluster terms taken from 10 documents and another two coders clustered terms taken from 5 documents. For the 10-document sets with more terms to sort, coders tended to form narrower or more specific categories compared to the broad categories formed for the 5 document sets. Unfortunately, the higher number of possible narrow topics means that different coders can choose different narrow topics to form clusters. This may explain the lower inter-coder similarity for the 10-document sets. On the other hand, lexical-based clusters are more similar between human coders and between machine and human coding!
Conclusion
The automatic method generated relatively good clusters compared to human clustering. According to the similarity measure used, the machine-generated clusters obtained a significantly higher similarity with the human-generated clusters (0.25) than the similarity between two human-generated clusters (0.19). However, the machine clustering did not take into account the semantics and thus cannot handle similar concepts having quite different syntactic forms, e.g. “computer”, “desktop” and “workstation”. Moreover, all the machine clusters were labelled with single-word terms that may have different meanings in different contexts. Thus, the clustered terms are sometimes not semantically very similar. For example, “computer network” and “social network” were assigned to the same cluster labelled “network”, though they are quite different in meaning.

The human coders were expected to use a semantic-conceptual approach to clustering terms. Instead, different coders were found to use different approaches, some using a combination of lexical and semantic approaches. Even when two coders use a semantic-conceptual approach, they often form different narrow categories since there are many possible narrow categories to “choose” from. Overall, this study has raised questions about how machine-generated clusters should be evaluated since there is so much variation in human clustering.

References
Cultural Impact on Knowledge Representation and Organization in a Subject Domain

Abstract
Culture affects the way in which knowledge is represented and organized. Cultural warrant in knowledge organization systems (KOS) can be considered as a quality indicator that would allow universal use. On the other hand, there is a demand for integration of cultural views in KOS in order to enrich the standard cultural and social perspective that these systems usually have. In order to allow cultural differences to be represented and organized in KOS in an integrated way, we must get to know those cultural differences so they can be represented and integrated to other views pertinent to the KOS, views on which it is intended to be built. This study aims to unveil how different cultures, as they are Spain and Uruguay, can impact in the developing of a subject field such as that of Gender Studies, and how this may affect to knowledge representation and organization in KOS. The final goal is to detect differences in terminology, categorization and conceptualization of the field studied, taking them as a reference point for suggesting an integrated and supranational proposal for knowledge representation and organization in KOS. To get this goal, an analysis of the domain has been conducted in order to know the dynamics and the content of Gender Studies in the two cultural areas studied.

Introduction
It has been claimed that culture affects the way in which knowledge is represented and organized. This is true even for subjects that are somehow considered to be more neutral as is the case with technology and science (Hassan, 2003). Cultural warrant in KOS (Beghtol, 2002) can be considered as a quality indicator that would allow universal use (Hunter & Beck, 2000; Treitler, 1996). On the other hand, there is a demand for integration of cultural views in knowledge organization systems (KOS) in order to enrich the standard cultural and social perspective that these systems usually have (Srinivasan, 2007; Rao, 2006; Kargbo, 2005; Muswazi, 2001; Liew, 2004 and Doyle, 2006; Zeng, Kronenberg & Molholt, 2001). In order to allow cultural differences to be represented and organized in KOS in an integrated way, we must get to know said differences so they can be integrated in the KOS that is intended to be built.

This paper aims to unveil 1) how different cultures, in this study Spain and Uruguay, can impact the development of a subject field, in this case the Gender Studies, that is much influenced by cultural and social environments, and to see how this fact is reflected in the specialized literature, 2) to show how this situation affects knowledge representation and organization in KOS as it is much based in literary warrant. To achieve this goal, we conducted an analysis of the domain, based on the terminology extracted from indexing of primary sources published in both countries, in order to detect the differences in terminology, categorization and conceptualization of the field of study, and 3) to make a comparison with the results found in these two cultural areas in order suggest an integrated and supranational proposal for knowledge representation and organization of the domain pertinent for both cultures.

Materials and Methods
The approach to domain analysis (Hjørland, 2002) of the Gender Studies has been the terminological one, in order to get a set of terms that represents this domain in both cultural areas. This was the starting point needed for a further study that aimed to get to know how to represent and organized the domain. To do so, the following procedure was carried out
for each area: 1) a study of how terms were deployed in order to observe the performance of those terms in the inter-discipline, and 2) a proposal for the conceptual organization of the Gender Studies domain based on the impact of the selected terminology as found in primary documents previously analyzed. Detailed data about the selected sources and about the methodology followed in the process of studying terms and conceptual structure for documents published in Uruguay can be found in López-Huertas, 2006. In the case of the Spanish publications, the selected sources consist of eight scientific journals until 2004, handbooks and electronic documents. The sources available in Spain represented important differences compared with those found in Uruguay, because there were no scientific journals on Gender published in Uruguay. On the contrary, articles about this matter were published in journals devoted to Social Sciences, conferences proceedings not specialized in Gender and chapters of books.

Indexing was done by assigning free key words. Terms were selected form titles, abstracts and main headings in monographs and articles. The obtained terminology was located in a relational data base on the Access platform that was designed for the study. It has seven fields: name of term, identification number, source of term provenance, source code, onomasiological variants, semasiological variants and context. This database can give results on request that are very helpful for this type of study, such as lists of terms, frequencies of terms, sources of the terms, etc. Terminology in the database was later examined according to a quantitative methodology based on the frequency of the terms’ appearance in the documents. The higher or the lower impact of the selected terms within the domain provided important information due to the fact that Gender is an interdisciplinary field. This procedure helped in the identification of different kinds of terms, many of them closely related to their original disciplinary provenance, and which showed how these “outsider” terms were incorporated into the interdisciplinary domain, according to the Gender epistemology.

**Results**

1) **URUGUAYAN AREA.** This region was studied in a previous paper that showed how the field of Gender Studies is represented in this context (López-Huertas, 2006). In this occasion, the resulting terminology (538 terms) was studied in two ways: 1) to discover the conceptual and terminological behaviour of this interdisciplinary domain and 2) to find out a pertinent conceptual structure for it.

The results showed that only a small proportion of the terminology is well represented in documents, but most of the terms are not, considering that the frequency is 1. A small amount of the terminology (32% of the total of descriptors) is found to be generated by the interdisciplinary activity itself, representing new organic matter. That is, it does not come from any other disciplines that interact within the domain. The rest of the terms (68%) are originated in other specialties, and it seems they are incorporated into the Gender domain apparently with the same form and meaning. Main subject areas forming the Gender domain were those dealing with the Social Sciences, Health/Hygiene, and Economy/Business, together with Gender itself, as shown in Figure 1. Figure 2 gives the thematic composition of main subject areas.

2) **SPANISH AREA.** Terminology extracted from indexing primary sources yielded a result of 1,421 terms. These terms were not well represented in documents since the 90.5% has only one hit in documents as seen in Figure 3.
From a knowledge organization point of view, we found noticeable differences between Spain and Uruguay. These results showed the necessity for the modification of some categories currently used. The results also showed the need for organizing sources in a different way from the conceptual structure suggested for the Uruguayan documents, in order to adjust the findings in the Spanish documents. Figure 4 shows the conceptual composition of the latter.

Quantitative and qualitative differences in knowledge representation were found between the two sets being compared. This fact effected, in turn, the suggested knowledge organization structure for each set. One of the most important differences was the impact of the Gender perspective on other specialties in the case of the Spanish documents. It was quite remarkable, not only because many documents dealt with a wide variety of disciplines, but also because new terms were created or stressed within outsider specialties due to the influence of the Gender perspective. This fact lead us to
create a new general category under the label *Gender perspective in Specialties*. In the case of the Uruguayan documents, a very small amount of documents on Specialties were found, 17 in total, that did not allow us to group them under a general category, instead an *Others* category was created to house them, as seen in Figure 2. As it can be seen in Figure 4, this group has 461 documents. An explanation of this big difference can be found in the fact that these studies are more developed in Spain than in Uruguay\(^1\). It seems that the maturity of the inter-discipline leads researchers in it to explore other knowledge domains close to their object of study. It is clear, by the data in Figure 4 that Gender Studies interacts most with Social Sciences and Humanities. It also shows that there is a remarkable social and cultural sensitivity towards this matter.

---

\(^1\) This claim is also based on the fact that the number of total terms (538) is much lower than the total of Spanish terms (1,421), along with the inexistence of journals specialized in Gender published in Uruguay.
Another interesting finding coming from Figure 4 was that the internal composition of main classes has changed, compared to the Uruguayan model. This change, in some instances has been so deep that it was needed to rename some categories as it is the case of Body/Mind, called Health/Hygiene in the Uruguayan knowledge domain. The main reason for doing that was that the group called Attitudes, Conducts, Feelings, Status was very important for the Spanish domain, and it was most representative of the interests of the Spanish scholars because it is the biggest group under this category with 108 hits. Nevertheless, it has almost no representation in the case of Uruguay. So, a new, more general name to label the category was needed. This group, in my view, is a very expressive matter for the Gender Studies domain, because many of the attitudes and feelings identified in the content of documents have been conditioned, in a way, by social constructions along with time.

It is also remarkable the weight that the groups Society/Social movements/Groups (200) and Politics/Policies (221) have within the class Social Environment. These matters are less important in the Uruguayan documents, where Politics is 34 and Society 13.

At the same time, the Uruguayan side presents thematic characteristics that are not pertinent for the Spanish domain. For instance, there is a concern about poverty or political and social situation in the area, as it can be seen in terms such as poor women, poor children, poverty, urban poverty, poor adolescents, family impoverishment, poor families, feminization of poverty, poor childhood, street children and street living, food discrimination, women execution, murders of women, missing women, country women, that
are not important for the Spanish area. Although it is not frequent, we found terms tied to Uruguayan traditions and customs such as charrúa women, murgas, murgueras, carnivals, carnival songs, Hispanic women, indigenous women, Latin-American women and Montevideanas.

Other points of thematic interest for Uruguay were concepts related to social networks to support women public and domestic activities, women's health and problems women face in labour market and job environments and discrimination. Figure 5 shows the comparison between both cultures where the said differences can be seen more clearly.

The comparison of main categories in both cultures in Figure 6 shows clear differences in the Gender Studies domain between the two cultures. In this Figure, we can see what has been said before in a clearer way. There is an evident difference of knowledge production between the two areas being compared that is materialized in a remarkable difference in the number of terms identified for each area. As it was said before, the main category Gender perspective in Specialties is almost non-existent in Uruguay. The class Body/Mind houses almost the group Health/Hygiene from the Uruguayan side but other groups in the same general category (Body/Image and Attitudes) are underrepresented.

Conclusions
After analysing the Gender Studies knowledge domain in two different cultures, even though they share the Spanish language, remarkable differences have been detected that take us to conclude the following:

1. Socio-cultural environments affect the thematic orientation of Specialties. An important part of the knowledge being produced is responding to socio-cultural demands, especially in subjects like Gender Studies. Culture influences the content of scientific (scholarly) production.
2. Differences in contents effect knowledge representation and organization that differ from culture to culture since KOS have to be built based on literary warrant.
3. In this paper, quantitative and qualitative differences have been detected in the production of specialized knowledge which affected common and shared conceptual structure, intended to be used by the two cultures, in two main ways: a) At a macro level, new main categories needed to be created when including the Spanish documents: Gender perspective in Specialties and Body/Mind. b) At a micro level, the importance of main categories based on the impact of terms in documents shows the need for deeper description or specificity. This way, the decision of how much to develop each class will be based in quantitative data coming directly from documents. At the same time, concepts directly linked to culture, as it was described above, can be accommodated in the structure giving to them the space for it.
4. The incursion of the Gender perspective on other specialties is interesting to observe because, in some instances, it created new core terminology inside other disciplines which means that Gender is somehow modifying the knowledge structures of those disciplines.
5. Cultural differences have to be borne in mind when planning to build KOS for use in multicultural environments. One way is to analyse the knowledge domain in each culture in a similar way as is described here. By doing this, documents will be indexed properly and users in each culture will find it familiar to their needs. We shall be building a system with cultural warrant as well.
References


A Pragmatic Perspective of E-mail Management Practices in Two Canadian Public Administrations

Abstract
This paper examines the prevalent contextual factors involved in the work of middle managers as they read and classify e-mail within two Canadian public administrations. Looking at e-mail through the lens of pragmatics and genre theory, the intent here is to devise a solution to alleviate issues associated with e-mail management practices. Resulting from this research, the E-mail Pragmatic Framework is presented. This framework takes into account prevailing individual, contextual and textual factors in the reading and classification of e-mail. As an alternative to speech act theory, a typology of e-mail pragmatic functions aiming to capture the expectations of middle managers as they interact with the e-mail genre is outlined.

1. Introduction
In the contemporary organisational context, e-mail is now entrenched as a critical communication medium. For large organizations, the burgeoning growth of the electronic message exchange is placing employees in a perpetual struggle to perform a vast amount of reading and classifying tasks often resulting in these tasks being postponed ad infinitum. Inspired by works that offer different perspectives to assist e-mail reading and classification, we conducted a study to better understand these practices in Canadian public administrations. The principal goal of this paper is to examine the prevalent contextual factors involved as employees read and classify e-mail. A second goal is to look at e-mail through the lens of pragmatics and genre theory to explore the suitability of such approaches in analyzing the contextual factors inherent to e-mail management. After reviewing the dominant works in these fields, we present the methodology of our own research and discuss the most salient results.

2. Literature Review
E-mail can be seen as a new genre (Goldstein and Sabin 2006). Genre is defined as a recurrent communication pattern that emerges from individual, social and technical forces creating expectations and facilitating the interpretation of information (Erikson 1999, 3). Prevalent work on genre theory in the organizational context is indebted to Joanne Yates and Wanda Orlikowski. These authors have elaborated the Genre Model (Yates and Orlikowski 1992, 2002; Orlikowski and Yates 1994; Yoshioka et al. 2001; Yates et al. 2008) as a tool to understand and evaluate “new technologies in the context of a specific organization, its goal, and its existing media usage” (Yates et al. 2008, 64). For the purpose of our study, the Genre Model provides interesting features in describing the context of e-mail usage with six key dimensions — why, what, who, where, when and how — which can be related to the field of pragmatics.

Why. This dimension deals with the purpose of the communication occurring in the genre and the users’ associated expectations (Yates et al. 2008, 64). The goal of a communicative genre is not only to transmit a meaningful message but also to suggest a specific course of action. To better understand the conventional actions linked to different genres in an organisational context, Yoshioka et al. (2001) suggest complementing the Yates and Orlikowski notion of purpose with speech act theory (Austin 1970; Searle 1972). Domi-
nant in the field of pragmatics, this theory stipulates that by saying something, a speaker performs three acts simultaneously: a locutionary act (what is said), an illocutionary act (what is done by saying it) and a perlocutionary act (the effects on the addressee). Given that e-mail genre is used, among other things, to negotiate and delegate tasks, several research endeavours propose using speech act theory to automatically classify e-mail based on the sender’s intention (Cohen et al. 2004; Carvalho and Cohen 2005, 2006; Goldstein and Sabin 2006; Khoussainov and Kushmerick 2007). These thinkers present different typologies of “e-mail acts” such as an amendment, request, commitment, proposals, greetings, and so on.

**What and How.** These dimensions deal with the form and content of a genre and the users’ associated expectations. The form refers to “readily observable physical and linguistic features” such as the structure, the communication medium and the symbol system (Orlikowski and Yates 1994, 544). Content refers to the “substance” of a genre defined as regular social motives, themes or topics (Yates and Orlikowski 1992). Research underpins the importance of content-related cues in the interpretation of the information as well as the mandatory articulation of the form versus content in the improvement of user access to digital text (Toms and Campbell 1999; Vaughan and Dillon 2006). Traditionally, research on filtering techniques has focused on the dominant e-mail skimming behaviours such as spam deleting, content summarization and keyword detection. Today, common messaging systems provide functionalities allowing users to create content or form-based rules to automatically classify their messages.

**Who and Where.** In the model, these dimensions deal with the social and the spatial dimensions of a genre and users’ associated expectations. These expectations vary in function of the participants and their roles as well as in the virtual or physical location of the exchange (Yates et al. 2008). In an organizational context, the use of genres can be associated with communities, whose members share “established practices, social relations, and communication media” (Yoshioka et al. 2001, 438). The question of who is involved in an e-mail exchange and the participants’ characteristics (work relations, frequency of communication, etc.) are important dimensions for e-mail interpretation (Dabbish et al. 2005). Recent studies have identified quantitative measures of social networks used to filter spam based on the personal e-mail network (Boykin and Roychowdhury 2004) or reputation ratings (Goldbeck and Hendler 2004).

**When.** This dimension deals with the temporal quality of a genre and the associated expectations as to timing and opportunity (Yoshioka et al. 2001, 437). Time is an important aspect in e-mail management (Gwizdka 2000). Methods have been developed to automatically extract time and date features in e-mail (Stern 2004) as well as to rank messages according to time-based priority indicators (Kadoya et al. 2004).

As demonstrated in the literature review, filtering technique research has focused on dominant e-mail skimming behaviours which are contingent on the presence of a genre. Even with these valuable results, there still remains a need for a more encompassing approach to articulate the 5W1H dimensions of the Genre Model to design a solution to better assist users with their e-mail management tasks.

### 3. Methodology

This qualitative study was undertaken in the context of two multidisciplinary research projects at the Université de Montréal, as well as a doctoral dissertation focusing on textual practices and genres in digital environments. The sample comprises eighteen middle
managers and eighteen administrative assistants selected in two Canadian public administrations. To obtain a broad perspective on how employees produce, use, and share texts in their work environments, thirty-six semi-directed interviews were conducted. To complement and validate these results, the same participants were asked to keep a diary journal during one regular working day. This journal recorded their textual practices describing the tasks involving paper and digital texts, the genres of texts solicited, the systems used, and the interactions with other people. Follow-up interviews were conducted the day subsequent to the diary completion.

It was immediately apparent that e-mail represents an important aspect of employee’s textual practices and that many problems encountered were due to an overload of messages. For a period of approximately forty-five minutes, five participants were asked to verbalize their thoughts while reading and classifying their e-mails. Vital information relating to the strategies used to skim through the messages and cues addressed while reading were revealed during this experiment.

4. Results and Discussion
This section exposes and discusses part of the results obtained during the data collection. First, the E-mail Pragmatic Framework is described. This framework accounts for individual, contextual and textual factors solicited while the middle managers and their administrative assistants — whose work entails e-mail management of their supervisors’ mail, from their supervisors’ perspective — read and classify e-mail. When identifying which cues can be used in a reading and classification solution, we found that speech act theory showed significant limitations. These limits are outlined in the second part of this paper. Finally, as an alternative to speech act theory, we present a typology of e-mail pragmatic functions aiming to capture the expectations of middle managers as they interact with the e-mail genre.

5. E-mail Pragmatic Framework
Figure 1 offers a bird’s-eye perspective of the E-mail Pragmatic Framework. Its development was guided by three rationales:

1. Within the circumstances of e-mail transmission, the framework prioritizes the receiver’s perspective, his/her motives and roles in the work context, and the strategies to determine the sender’s communicative intentions. The relation, depicted
by the arrow between the cognitive, the contextual and the textual strata, shows the receiver’s perspective as the locus of interpretation.

2. User interaction with an e-mail corresponds to a linguistic act. E-mail is defined as a genre comprising two levels: (1) an interface level, determined by the messaging system and (2) a textual level. Therefore, interaction is envisaged following its physical (the manipulation of the interface), cognitive (the comprehension of the content) and discursive (the interpretation of language) dimensions.

3. The framework takes into account textual as well as non-textual (implicit) cues in the interpretation of information.

The data have revealed that e-mail management practices are affected by two main cognitive triggers, namely the priority and the implication levels assigned to each message. These results highlight the relevance of adopting a pragmatic perspective in the study of e-mail as a general theory of human action (Maingueneau 2005, 3). The process of determining the level of priority and the action to be performed is supported by several contextual factors relating to the six dimensions of the Genre Model (Yates et al. 2008). These include the timeframe (when), the sender name (who), the action and its position in the conversation (how), the object of the action, and the topic or case (what, where, if applicable). These contextual factors are, in turn, assessed through a series of textual features (how), as well as implicit cues articulated by the reader.

The results also confirmed the value of speech act theory for further study of e-mail management practices. Participants effectively classified their e-mail according to broad categories of actions. For these action categories, common objects were repeated such as “attending a meeting”, “confirming a meeting”, “cancelling a meeting”, “planning for a meeting”, etc. But this perspective, inspired from works on “e-mail speech acts” and their objects (Cohen et al. 2004; Scerri et al. 2007), is insufficient to tackle the complexity of e-mail interaction. The process of identifying textual cues to design an e-mail reading and classification aid leads us to reconsider the use of speech act theory. In the next section, we briefly outline three significant limitations.

**Limitations of Speech Act Theory in Studying E-mail Management Practices**

The first limitation in this application results from the strict translation of Searle’s classification of illocutionary acts to the study of e-mail. According to speech act theory, the illocutionary act corresponds to the act of doing by saying something as opposed to the perlocutionary act which corresponds to the illocutionary act effect on the addressee. However, the boundary is blurred when determining the actual illocutionary and perlocutionary values of a linguistic act (Kerbrat-Orecchioni 2005, 22). If we adopt an approach that distinguishes the illocutionary intent from the perlocutionary effect, we realize that authors who use Searle’s illocutionary acts to classify e-mail messages do not take into account their effects on the recipients. However, in our study, the implication (Do I have to do something after reading this e-mail?) is what is of key importance to the participants: it is therefore the perlocutionary dimension of the illocutionary act itself that is of value in an e-mail reading and classification aid.

The second limitation resides in the fact that Austin and Searle’s speech acts appear to be “abstract and isolated entities, detached from their actualizing context and from other acts that may precede and follow in a discursive sequence” (Kerbrat-Orecchioni 2005, 53, our translation). For e-mail study, this limitation presents serious impacts. E-mail generates typical action/reaction sequences which are significant for the recipient.
In our research, the *position* of a message in the thread of a conversation was important for the participants (*Does this message answer one of my requests?*). In addition, the thread sequence cannot always be considered a direct exchange (sender X to recipient Y to sender X), as e-mail is often multi-voiced. In the context of an exchange, recipient Z, for instance, can receive messages in copy (CC) retaining traces of the original exchange between X and Y. Even if the actions conveyed in these messages are the same, the interpretation and resulting action would be different.

The third limitation resides in the difficulty of identifying speech act units of analysis. If the various speech acts are considered part of a broader conversation, clarity is then needed to determine their extent (in the sentence, the paragraph, or the whole text) (Kerbrat-Orecchioni 2005, 54). Some research suggests referring to the “macro speech act” concept similar to that of genre to account for a pragmatic unit exceeding the sentence (Maingueneau 2005, 12). Orlikowski and Yates (1994) underpin this distinction by detailing the repertoire of electronic mail into four genres: the memo, the dialogue, the proposal and the ballot. We believe e-mail is an emerging genre that could be subdivided into subgenres, following common characteristics of form, content and purpose. Nevertheless, the challenge remains in the necessity to enable speech act theory to better support e-mail users while considering other important contextual factors.

**Typology of Prevalent E-mail Pragmatic Functions**

Inspired from the *E-mail Pragmatic Framework*, a typology of middle managers’ e-mail pragmatic functions was developed (Figure 2).

The structure aims to facilitate the evaluation of the recipient’s expected implication upon reading a message and the assignment of a priority level. The *action* category regroups e-mail pertaining to Searle’s directive illocutionary acts such as a question or a command. The *reaction* category shows the response of the sender to a demand of the recipient. The *direct follow-up* gathers the “CC” and the informational e-mail sent by a person pertaining to the recipient’s social group. The *indirect follow-up* category groups

<table>
<thead>
<tr>
<th>E-Mail Pragmatic Functions</th>
<th>ACTION</th>
<th>REACTION</th>
<th>DIRECT FOLLOW-UP</th>
<th>INDIRECT FOLLOW-UP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Giving an AUTHORIZATION</td>
<td>Giving an ANSWER or an OPINION/ Transmitting a DOCUMENT or an INFORMATION</td>
<td>Promising or confirming the completion of a TASK</td>
<td>Release / Advertising</td>
</tr>
<tr>
<td></td>
<td>Performing a TASK</td>
<td>Confirming or deciding a MEETING</td>
<td>Reminders (SENDER = RECIPIENT)</td>
<td>Untargeted information</td>
</tr>
<tr>
<td></td>
<td>Participating to a MEETING</td>
<td></td>
<td>Targeted information (MANAGER team and SUPERVISORS)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ACTION</th>
<th>REACTION</th>
<th>DIRECT FOLLOW-UP</th>
<th>INDIRECT FOLLOW-UP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Giving an AUTHORIZATION</td>
<td>Giving an ANSWER or an OPINION/ Transmitting a DOCUMENT or an INFORMATION</td>
<td>Promising or confirming the completion of a TASK</td>
<td>Release / Advertising</td>
</tr>
<tr>
<td>Performing a TASK</td>
<td>Confirming or deciding a MEETING</td>
<td>Reminders (SENDER = RECIPIENT)</td>
<td>Untargeted information</td>
</tr>
<tr>
<td>Participating to a MEETING</td>
<td></td>
<td>Targeted information (MANAGER team and SUPERVISORS)</td>
<td></td>
</tr>
</tbody>
</table>

In our research, the *position* of a message in the thread of a conversation was important for the participants (*Does this message answer one of my requests?*). In addition, the thread sequence cannot always be considered a direct exchange (sender X to recipient Y to sender X), as e-mail is often multi-voiced. In the context of an exchange, recipient Z, for instance, can receive messages in copy (CC) retaining traces of the original exchange between X and Y. Even if the actions conveyed in these messages are the same, the interpretation and resulting action would be different.

The third limitation resides in the difficulty of identifying speech act units of analysis. If the various speech acts are considered part of a broader conversation, clarity is then needed to determine their extent (in the sentence, the paragraph, or the whole text) (Kerbrat-Orecchioni 2005, 54). Some research suggests referring to the “macro speech act” concept similar to that of genre to account for a pragmatic unit exceeding the sentence (Maingueneau 2005, 12). Orlikowski and Yates (1994) underpin this distinction by detailing the repertoire of electronic mail into four genres: the memo, the dialogue, the proposal and the ballot. We believe e-mail is an emerging genre that could be subdivided into subgenres, following common characteristics of form, content and purpose. Nevertheless, the challenge remains in the necessity to enable speech act theory to better support e-mail users while considering other important contextual factors.

**Typology of Prevalent E-mail Pragmatic Functions**

Inspired from the *E-mail Pragmatic Framework*, a typology of middle managers’ e-mail pragmatic functions was developed (Figure 2).

The structure aims to facilitate the evaluation of the recipient’s expected implication upon reading a message and the assignment of a priority level. The *action* category regroups e-mail pertaining to Searle’s directive illocutionary acts such as a question or a command. The *reaction* category shows the response of the sender to a demand of the recipient. The *direct follow-up* gathers the “CC” and the informational e-mail sent by a person pertaining to the recipient’s social group. The *indirect follow-up* category groups
the untargeted information, the releases, and advertisings. These two last categories present a low level of implication for the recipient.

The structure integrates the conversational dimension pertaining to the e-mail genre with the creation of categories such as action/reaction. These categories aim to reflect the exchange dynamics between a sender and a recipient while keeping track of a message in a conversational thread (such as in Scerri et al. 2007). As mentioned earlier, this aspect is important as it varies the perlocutionary effect of the message from an action to perform to a simple consultation of information. The structure also distinguishes the actions from their objects. As the analysis of single action verbs is not enough to capture the recurrence of one speech act’s propositional content (Winograd and Flores 1987, 161), the addition of objects such as authorization or meeting allow for a richer presentation of e-mail content.

Finally, the typology also supports the modeling of the recipient’s social network. For example, the sub-categories direct follow-up versus indirect follow-up take into account the hierarchical position of the sender’s role in relation to the recipient. In our study, the participants gave great importance to the sender’s name when assigning a priority level to the messages (also in Dabbish et al. 2005).

6. Conclusion

By gaining better insights into the contextual factors pertaining to e-mail management practices, this research presents several theoretical and practical implications. In the context of Canadian public administrations, this research leads to the possibility of a tailored typology of recurring forms of e-mail actions. This typology considers more than speech act theory and incorporates several important dimensions of users’ expectations while interacting with the e-mail genre. Further steps entail the application of the e-mail pragmatic functions typology to the analysis of a corpus of ca. 1600 French messages received by two middle managers. The objective is to devise a solution to alleviate issues associated with e-mail reading and classification tasks while still accounting for contextual factors inherent to e-mail management practices.

Acknowledgement

This research is kindly supported by research grants from Fonds québécois de recherche sur la société et la culture and Chaire Bell en recherche interdisciplinaire sur les technologies émergentes as well as a doctoral scholarship to Inge Alberts from Social Sciences and Humanities Research Council of Canada.

References


Daddy, How do I Find a Book on Purple Frogs? Representation Issues for Children and Youth

Abstract
Subject access and controlled vocabularies used for representing children’s resources are examined. Wittgenstein’s Language Games theory is presented as a possible framework for determining sources to use for controlled vocabulary construction. Preliminary results regarding existing controlled vocabulary use in databases and digital libraries designed for children are presented, as are extent of match findings comparing users’ search terms with LCSH, Sears, and a subject-specific proprietary controlled vocabulary are reported.

Introduction
A key challenge to retrieval in any type of system is how to represent the resources appropriately so that the user(s) can find what they are looking for. In systems being used by or designed for children and youth\(^1\), there exist two fundamental representation problems: (1) the metadata or representation scheme of the system may not be designed with this specific user group in mind, and (2) few age-appropriate controlled vocabularies exist for use in creating metadata.

Research has shown that children and youth have unique information needs and information seeking strategies (for an overview see Abbas (2005a); Abbas, in press; Walter (1994)). In today’s increasingly digital world, children have access to a variety of global resources in many different formats. Children have access via OPACs in libraries, proprietary databases, and specialized collections of resources in digital libraries, subject directories and web portals that are designed specifically for their use. As the Web continues to develop and more Web 2.0 applications are introduced, children will access information in many differing contexts and formats, for example, social sharing sites like YouTube or MySpace where they can search through dynamic content and formats like videos, photos, blogs and wikis. Each of these different systems and information ecologies are organized by different underlying structures and standards, or lack thereof.

In order to maximize the success of children’s information seeking, the system design and the resources contained within the system must be represented at a level appropriate for this particular group of users. The metadata scheme and the metadata describing the system resources should reflect an understanding of how children access, organize, and use information (Abbas, 2005a; Abbas, 2005b). In social sharing sites, users control the content and the basic structures, though they design their “spaces” using pre-determined sets of available features provided by system designers/programmers.

Representation of resources is a “complex web of attributes of disparate objects and concepts, idiosyncratic and socially constructed codes and agreements, and neurological abilities” (O’Connor, 1996 p. 11) and has been defined using many lenses such as library and information science, cognitive science, and linguistics, among others. Representation for children is a complex socio-cognitive process in which many variables come into play: 1) user’s developmental and cognitive state, 2) user’s domain and system knowledge, 3) indexer’s knowledge of the user’s information seeking habits, 4) indexer’s understanding of the intended purpose(s) for the objects, and 5) the controlled vocabu-

---
\(^1\) For the ease of the reader, the term children will be used throughout to represent all younger users between ages 0–15.
lary used to create representations. The representation process is further complicated by: 1) our incomplete picture of this group of users information seeking and system use, 2) metadata schemes designed for use by adults and not children, 3) the guidelines and tools to construct metadata, and 4) the differing cognitive abilities and system knowledge of children.

Blair (1990, 2006) sees the problem of representation and information retrieval as linguistic in nature. How effectively we utilize language to represent an object, determines the success or failure of the information retrieval process. Blair also posits that the language that we use to represent both our information needs or questions and that are used to index documents is learned in a social context or community. Blair explains the theory of “language games”, as first developed by the early twenty century philosopher Ludwig Wittgenstein, and the process in which we learn language and meaning. We do not acquire language purely by learning the word and its definition, but instead learn its use and appropriateness within the context of our “forms of life” or everyday experiences. Children engage in “language games” as they go through their daily “forms of life” or interactions. Direct influences on their learning are their parents, teachers, the documents with which they engage (books and textbooks, web resources, popular and trade media in many formats such as magazines, movies, social sharing sites (YouTube, MySpace, etc.)), and the information systems they interact with. Learning the appropriate “language” or terminology to use within these contexts is vital to their success both in information retrieval and content understanding.

The tools provided the cataloger/metadata creator obviously also can affect successful retrieval by children. A cataloger must have an age-appropriate controlled vocabulary to use for representing subject content of children’s resources. The most prevalent controlled vocabulary in use in library systems, the Library of Congress Subject Headings (LCSH) contains words that are at mostly above a sixth grade level, thereby making their use inappropriate for younger users. The Annotated Card (AC) Program of the Library of Congress is a subset of LCSH that has been altered to fit more appropriately for use in cataloging children’s resources. The AC also provides guidelines to follow to standardize cataloging for children. The Sears List of Subject Headings is another controlled vocabulary that can be used to create metadata for children’s resources.

Few researchers have explored the issue of developing more age-appropriate controlled vocabularies for children. Abbas (2005b) explored the idea of using children’s frequently used search terms as a source for controlled vocabulary terms. Jannson (as reported by Lundgren, 1998) developed a special thesaurus for children consisting of about 800 simple, concrete words within 21 areas of interest. Pejtersen’s Book House project developed an icon-based interface for an OPAC. What is most interesting about the Book House project is that it does not use a traditional metadata scheme such as MARC. The bibliographic records include additional information such as the emotional effect the book may produce and descriptions and evaluative comments written by 11 and 12 year old children. (Lundgren, 1998)

Of particular interest in the study being reported here is the application of Wittgenstein’s “Language Games” (LG) as a means to develop an age-appropriate controlled vocabulary for children and youth. Using LG as a framework, the study examined: 1) existing controlled vocabularies used to catalog children’s resources; 2) the usefulness of existing controlled vocabularies, 3) identified a set of sources present in LG “forms of life” encountered by children as they learn to use language for interacting and searching;
and 4) compared existing controlled vocabularies with the terminology used within LG sources.

**Methodology**

**Research Questions**

To guide this study a set of research questions were formulated:

1. Which controlled vocabularies are used by catalogers of children’s resources?
2. Which controlled vocabularies are used by systems used by/designated for children?
3. Is Wittgenstein’s “Language Games” (LG) theory a useful framework for determining sources to construct a controlled vocabulary for children?
4. Using LG theory, which resources can inform controlled vocabulary development?
   a. Do users’ search terms match controlled vocabulary terms used by the system?
   b. Do users’ search terms match terms used in LG sources?

This paper will report preliminary findings of questions 1, 2, 4, 4a and 4b. The exploration to date serves as a “proof of concept” for examining question 3.

**RQ 1:** Standards for controlled vocabulary use in library catalogs are determined by the library itself, the country’s library governance body, and by the discipline. For example, in the U.S., the Association for Library Collections & Technical Services (ALCTS), Cataloging and Classification Section, Cataloging of Children’s Materials Committee, division of the American Library Association developed a set of guidelines catalogers can follow when cataloging children’s resources. These guidelines establish principles for applying Library of Congress’s Annotated Card (AC) Program headings. They also provide guidance for using other age-appropriate controlled vocabularies, such as Sears List of Subject Headings, or curriculum-specific headings. The guidelines were developed to solve the issue of non-standardized cataloging of children’s resources and to dissuade catalogers who contribute to shared cataloging from adding local subject headings into their records. (ALCTS; Inttner, Fountain, & Gilchrist, 2006)

Unfortunately no complete answer to this research question could be determined by reviewing the literature related to children’s cataloging. Much of the literature is anecdotal in nature and discusses current or best practices or is focused on subject access in general, not specifically for children. Weihs (2004) reports on her “unscientific survey” of the use of Sears List of Subject Headings. She emailed friends and contacts given by friends in the U.S. and Canada (no sampling frame or size given), and asked them: 1) if they knew the difference between Sears and LCSH; and 2) whether or not they used Sears List of Subject Headings or some other controlled vocabulary. The results revealed: 1) most school librarians did not know which controlled vocabulary was used in their library catalog, 2) many used CIP (Cataloging in Publication) information only, 3) some reported not knowing the difference between Sears and LCSH; others thought LCSH was the only controlled vocabulary; and 4) the librarian does not always make the decision regarding the source of cataloging information, this is often determined by the school board. Weihs also notes a personal communication with the editor of Sears, Joseph Miller, who reported that Sears is used “in countries where there is some library collection or education in English” (2004, p. 6). While the Weihs “unscientific survey” reveals a bit of the picture of U.S. and Canadian understanding and use of Sears and LCSH, it is really not a definitive answer to the research question posed.
Other useful international efforts related to subject access issues in general have been conducted or are in process. These studies, however, do not focus on subject access for children’s works. Heiner-Freiling (2000) presents findings of an IFLA Section on Classification and Indexing survey. The goal of the study was to determine subject heading languages used in national bibliographies or catalogs of national libraries. Responses from 88 national libraries or national bibliographic agencies were received. The responses were from 37 European countries; 18 African countries; 15 Asian countries; 15 countries in the Americas; and from 3 other countries (Australia, New Zealand, Papua New Guinea). The results show: 1) twenty-four national libraries use LCSH and twelve countries use translations or adaptations of LCSH; 2) other subject heading lists including RAMEAU and the German subject heading system RSWK (Regeln für den Schlagwortschatz) are gaining prominence in the countries sharing French or German as the common language; and 3) Sears List of Subject Headings as a second English language subject heading system was also shown. It is being used in ten countries, mostly those with English as the first or second language (e.g., the Bermuda Islands, Belize, Barbados, Ethiopia, Zambia, and Bangladesh) and it is being used for cataloging children’s works (Heiner-Freiling, 2000, pp. 191–192). Landry (2007) reports on efforts currently underway by the IFLA Working Group on Guidelines for Subject Access by National Bibliographic Agencies to establish guidelines for subject access by national bibliographic agencies and to develop an indexing policy for use in national bibliographic catalogs. The work of this group may give us more of a complete picture of controlled vocabularies used internationally for subject access, however, it is not noted whether it will examine those used specifically for children’s resources.

Next steps include: 1) review IFLA findings when available, and 2) develop and conduct survey of school and public librarians to determine controlled vocabularies being used.

**RQ 2:** Controlled vocabulary use in online and commercial systems varies. It is difficult to determine which vocabulary a system is using because of the limited documentation on the systems’ web sites and/or the lack of literature containing this information. In many cases Commercial online databases such as Gale Group, EBSCOHost and Proquest use content specific metadata schemes and subject specific controlled vocabularies, such as the ERIC thesauri for educational resources in the ERIC database. Web metadata creators use either generic metadata schemes such as the Dublin Core (DC) to represent resources or they adapt existing metadata schemes such as MARC and DC by adding additional elements (or database fields) that are subject or audience specific. Additionally, specialized metadata schemes are being designed with a particular group of user(s), resources, and use in mind. For example the Gateway to Educational Materials (GEM) metadata scheme has been developed using the DC as a base set of elements. Additional elements were then added to make the scheme more useful to the users of the collection, K-12 educators. Multiple communities with specific users and needs are currently developing metadata schemes for systems and resources designed for adults but few have children as their focus. (Abbas 2005a)

To explore this question, a set of online commercial databases (Ebsco’s Science Reference Center, and Ebsco’s General Science Collection) and one national digital library (National Science Digital Library) NSDL were reviewed to determine which controlled vocabularies were used to represent resources for children. The research at this point
chose to focus on science-related systems in order to narrow the subject areas being explored. In future explorations, other subject areas will be included as well. The following steps were conducted:

1. Emails and phone calls to representatives of the companies were sent but none returned to date.

2. Each site was reviewed for information about the controlled vocabulary used. The proprietary databases included FAQ and search help pages. The site explained the subject authority file as “a controlled vocabulary” [being] “drawn from the subjects assigned to each record” which would lead one to assume it is a controlled vocabulary constructed by terms derived from the resources themselves (natural language). However this assumption could not be verified. When the NSDL was reviewed, it was noted that multiple subject-specific controlled vocabularies are in use. Due to the number of contributors to the NSDL and the scope of the collections included, NSDL determined that it is not feasible to have one all encompassing controlled vocabulary for use in all collections. The Library of Congress Subject Headings was noted but contributors to NSDL continue to use their own controlled vocabularies. NSDL instead provides cataloging best practices for contributors to follow. (NSDL)

3. A sample of 33 search strings by 5th–8th grade users were run in the Ebsco General Science Collection online database. These same terms were also compared to the Library of Congress Subject Headings list (both standard list and Annotated Card) and to Sears List of Subject Headings. The findings are presented in Table 1.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>EM</td>
<td>21 (63%)</td>
<td>2 (6%)</td>
<td>21 (63.6%)</td>
<td>22 (66.6%)</td>
</tr>
<tr>
<td>Derived</td>
<td>5 (15%)</td>
<td>0</td>
<td>2 (6%)</td>
<td>5 (15%)</td>
</tr>
<tr>
<td>Paren. Qual.</td>
<td>3 (9%)</td>
<td>0</td>
<td>2 (6%)</td>
<td>2 (6%)</td>
</tr>
<tr>
<td>USE ref.</td>
<td>2 (6%)</td>
<td>1 (3%)</td>
<td>1 (3%)</td>
<td>1 (3%)</td>
</tr>
<tr>
<td>No term</td>
<td>1 (3%)</td>
<td>30 (90%)</td>
<td>1 (3%)</td>
<td>1 (3%)</td>
</tr>
</tbody>
</table>

It is interesting to note that in all controlled vocabularies used by the systems, with the exception of AC, the majority of the users’ search terms exactly matched those in the controlled vocabulary and only 1 (3%) of the terms did not match any of the controlled vocabularies. In other cases, the search term matched with a derived term in the controlled vocabulary (plural form or extended form) or contained a parenthetical qualifier. Few of the terms (between 3–6%) were unauthorized terms. These findings indicate that for this sample of terms, the controlled vocabulary of the system actually matches to a good degree with those terms used by the searchers. This is an important and positive finding for the existing controlled vocabularies being used by systems (LCSH, Sears, and Ebsco).

2. The terms used were chosen from a sample of searches run over one school semester in the GoKnow digital library. The terms were the most frequently used search terms used by this set of users, having been used in searches in a range of 50–519 times. For a list of the terms and table of results, please email author at abbasjm@buffalo.edu.

3. While the print and online versions of the AC did not include many of the terms in the sample, AC guidelines have provisions for either using existing LCSH terms or for adding terms.
Next steps will include: 1) expanding the sample of terms and systems to include non-science related terms, 2) run the queries in the non-science related systems, and 3) follow up with all database vendors for more information on controlled vocabularies.

RQ 4: Identifying LG sources to use is the heart of the further study. As mentioned above children encounter many sources in their daily “forms of life” that could inform us about controlled vocabulary development. This preliminary study examined one of those sources, science textbooks used by 8th graders, but further exploration is warranted and more potential sources need to be identified and reviewed. To test the efficacy of this idea, one 8th grade science textbook (used by students ages approximately 10–12) was reviewed. The glossary, index, text of the chapters, and chapter review word banks were examined to determine if the search terms used in RQ #3 appeared as prominent terms in the textbook. Of the 33 search terms, 14 or 42% of the terms appeared in the textbooks in one or many of the areas: 1) 6 terms or 18% were in glossary entries; 2) 12 terms or 36% were in index entries; 3) 4 terms or 12% were in word banks; and the sample terms appeared prominently 65 times in the textbook.

Next steps will include: 1) expanding this idea to more textbooks and new set of non-science terms; and 2) identifying and examining other LG sources to use.

Conclusion
The study presents some intriguing findings but also shows areas that need further exploration. Each of the research questions proves to be useful focus areas to explore the issue of children’s representation further. The Language Game theory as a framework for the development of a more-age appropriate controlled vocabulary was shown to be useful, though more sources need to be identified and examined. While this study adds to the knowledge we have regarding subject access in systems designed specifically for use by children and the use of existing controlled vocabularies to catalog children’s resources within databases and OPACs, it also shows the large gaps in the research and literature concerned with this issue.

References


Ethics in the Knowledge Organization Environment
An Overview of Values and Problems in the LIS Literature

Abstract
Reflections on Information Science have been focused mostly on information access and dissemination, not on ethical aspects of knowledge organization and representation (KOR). This leads us to investigate the existence of ethical values — and problems — which have impact on this field, especially since they are not discussed, although they are revealed in everyday practice. Therefore, and trying to contribute to a further reflection on the lack of literature on ethics in KOR, this paper analyses the Journal of the American Society for Information Science and Technology — JASIST, Journal of Documentation, Knowledge Organization, Cataloging & Classification Quarterly, The Indexer and Ethics and Information Technology in the period between 1995 and 2004. The results reveal two complementary dimensions: one concerning the respect of diversity and the other concerning the specificity warrant. The latter, which may prove the relevance of the theoretical principles announced by Hudon (1997), Beghtol (2002, 2005) and García Gutiérrez (2002), relative to a transcultural ethics of mediation that reflects diversity (fitted with a precision that, many times, passes by the dimension of multilingualism), making use of tools that may provide cultural warrant to knowledge representation.

Introduction
Starting from the philosophical conceptualization of ethics as a discipline which studies the values and the ways man conducts himself within society, especially concerned with the rules of conduct which must lead rights and duties in the social context (Rosental & Iudin, 1959), we have come to the conceptualizations of welfare, right action, duty, obligation, virtue, liberty, rationality, choice and also to the subjective features involved in each of these concepts (Blackburn, 1997). So, by its nature, ethics can be located within the scope of a normative knowledge, a guide on which human conduct should be based, yet allowing reflective thinking on different morals and different ways of rationally justifying moral life (Cortina & Martínez, 2005).

In this sense, special attention should be given to Kant’s notion of *categorical imperatives*, as actions objectively important and necessary by themselves, without relation to any other end (Kant, 1994, 124–125), which relate to the idea of values, not only those considered universal and impartial but also those which derive from moral ideas and cultural roots and survive under the lights of reason. In this way, ethical values may be viewed as beliefs whose aim is to achieve what is conventionally denominated as ideal of right action, while human behaviors which do not follow or respect these beliefs may reveal ethical problems.

In professional contexts, debates concerning ethical values are of special relevance and involve the so called Professional Ethics or Deontology whose object of study is the set of moral values a specific profession must be guided by, in order to achieve a professional attitude and action which must be correct and suitable to the social context in which it is inserted. It is common to have this set of guiding values of a specific profession made visible to the target community, disseminating it by means of rules clearly expressed in ethical codes.
The international literature on Knowledge Organization (KO) within Information Science context, presents some ethical concerns, mostly related to the prejudices and antipathies in indexing languages (Greenblatt, 1990), to the possible indexer’s negligence on the moment of choosing indexing terms (Dahlberg, 1992), or to the enormous power the indexer is granted on choosing terms during the subject analysis (Olson, 2002) as well as to the loss (and even damages) suffered by an author when his/her work cannot be compiled by an international publication due to insufficient, inept or biased indexing (Van der Walt, 2002).

Leblond (1999) and Arot (2000) refer to the commitment of the librarian to the system spirit, revealing professional values of information inter-relationship, ordering and access. Guimarães (2000) proposes four levels of ethical commitments — the user, the organization, the information and the profession. Beghtol (2002, 2005) presents the notions of cultural warrant and cultural hospitality which can establish dialogical relations with the concept of transcultural ethics in information mediation proposed by Garcia Gutiérrez (2002) as the first step to the establishment of a borderline culture which can then serve as context and/or reference to account for the complex and hybrid nature of transcultural relations. Such an approach is compatible with the concerns of Hudon (1997), especially on issues related to conceptual and terminological structures of each language in multilingual thesauri, assuring respect of difference in such a way as to provide each of the target user communities with the necessary terminological familiarity to them. In short, all this reveals respect to the political and cultural dimensions involved in the process and assures quality and equalitarian access to information (Pinho, 2006).

It is important to call attention to the concern of Thellefsen & Thellefsen (2004) on the need of indexing being committed to the structure and the reasoning of knowledge areas; also to point out the position of Bair (2005) who views information retrieval by the user as a fundamental requirement to the preservation of a greater ethical value: access to information.

To verify how this theoretical-conceptual scenario is displayed in the specialized periodical literature of KO, analysis of some of the most important international specialized journals was carried out over a period of ten years. The starting point was the presence of some specific terms in the title, in the abstract and in the key words, a methodology which had previously been used by Guimarães & Fernández-Molina (2003) while analyzing the periodical Knowledge Organization. The publications selected for the study were: Journal of the American Society for Information Science and Technology — JASIST, Journal of Documentation, Knowledge Organization, Cataloging & Classification Quarterly, The Indexer, and Ethics and Information Technology which were analyzed from 1995 to 2004. Considering that JASIST and Journal Documentation refer to a broader range of subjects, they were analyzed under two conceptual domains: ethics (represented by the terms Ethics; Ethical; Ethos; Deontology; Value(s); Conduct; Moral) and knowledge organization (represented by the terms Knowledge Organization; Knowledge Representation; Indexing; Classification; Subject Cataloguing; Subject Analysis; Content Analysis; Abstracting; Thesaurus(i) and Subject Headings). Cataloging & Classification Quarterly and The Indexer were analyzed under the domain ethics since they present a more specialized approach in KOR.
Results and discussion
The analysis of the results allowed us to identify a group of 23 articles whose contents were related to ethical questions in KOR (the occurrences of the ethical questions are represented by the numbers in Table 1). In this sense, it was possible to identify a set of ethical values as follows:

<table>
<thead>
<tr>
<th>Table 1. Ethical values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Values</strong></td>
</tr>
<tr>
<td>Privacy</td>
</tr>
<tr>
<td>Precision</td>
</tr>
<tr>
<td>Cultural warrant</td>
</tr>
<tr>
<td>Autorship (copyright)</td>
</tr>
<tr>
<td>Exhaustivity</td>
</tr>
<tr>
<td>Competence skill</td>
</tr>
<tr>
<td>Consistency</td>
</tr>
<tr>
<td>Accessibility</td>
</tr>
<tr>
<td>Efficiency</td>
</tr>
<tr>
<td>Flexibility</td>
</tr>
<tr>
<td>Freedom</td>
</tr>
<tr>
<td>Information safety</td>
</tr>
<tr>
<td>Reliability</td>
</tr>
<tr>
<td>Equity</td>
</tr>
<tr>
<td>Usability</td>
</tr>
<tr>
<td>Hospitality</td>
</tr>
<tr>
<td>Professional recognition</td>
</tr>
<tr>
<td>Up-to-dating</td>
</tr>
<tr>
<td>Autonomy</td>
</tr>
<tr>
<td>Power awareness</td>
</tr>
<tr>
<td>Co-operation</td>
</tr>
<tr>
<td>Information diversity</td>
</tr>
<tr>
<td>Risk minimization</td>
</tr>
</tbody>
</table>

The mentioned values can be grouped in three spheres:

a. **superior values which must guide all the informational activity**, such as Respect to Privacy, Authorship (copyright), Accessibility, Freedom, Safety, Equity, Diversity and Risk Minimization;

b. **values previously recognized as professional requirements because they are essential to the action of an information professional**, such as Competence-Skill, Efficiency, Flexibility, Reliability, Professional Recognition, Up-to-dating, Autonomy, Power Awareness and Cooperation;

c. **values previously considered as mere information retrieval measures, but recognized today as part of the axiological universe of KOR**, such as Precision; Recall; Cultural Warrant, Exhaustivity, Consistency, Usability, Hospitality.

In this context, five fundamental values, which together correspond to 57% of the whole analyzed universe, have been observed: Privacy, Precision, Cultural Warrant, Authorship (copyright) and Exhaustivity.
It was also possible to identify and systematize the following ethical problems in KOR, which occur mainly for the refusal or negation of the values (the occurrences of the ethical problems are represented by the numbers in Table 2):

<table>
<thead>
<tr>
<th>Problems</th>
<th>JASIST</th>
<th>J. Doc.</th>
<th>C&amp;CQ</th>
<th>The Indexer</th>
<th>EIT</th>
<th>TOTAL (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surveillance</td>
<td>12</td>
<td></td>
<td>12</td>
<td></td>
<td></td>
<td>21.82</td>
</tr>
<tr>
<td>Censorship</td>
<td>4</td>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td>7.27</td>
</tr>
<tr>
<td>Digital segregation</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>7.27</td>
</tr>
<tr>
<td>Lack of cultural warrant</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td>7.27</td>
</tr>
<tr>
<td>Negligence</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.27</td>
</tr>
<tr>
<td>Informational Directness</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>5.45</td>
</tr>
<tr>
<td>Professional inefficiency</td>
<td>2</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>5.45</td>
</tr>
<tr>
<td>Misrepresentation</td>
<td>2</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>5.45</td>
</tr>
<tr>
<td>Racism</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>5.45</td>
</tr>
<tr>
<td>Ambiguity</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>3.64</td>
</tr>
<tr>
<td>Marginalization</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.64</td>
</tr>
<tr>
<td>Pornography</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td>3.64</td>
</tr>
<tr>
<td>Impartiality or neutrality belief</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.82</td>
</tr>
<tr>
<td>Defamation</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.82</td>
</tr>
<tr>
<td>Discharge of electronic garbage</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.82</td>
</tr>
<tr>
<td>Idiosyncrasy</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.82</td>
</tr>
<tr>
<td>Inaccessibility to information</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.82</td>
</tr>
<tr>
<td>Professionals being replaced by technology</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.82</td>
</tr>
<tr>
<td>Prejudices in terminology</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.82</td>
</tr>
<tr>
<td>Biased translations</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.82</td>
</tr>
<tr>
<td>Violence</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>1.82</td>
</tr>
</tbody>
</table>

By the same methodological procedure, it was possible to categorize the ethical problems in two universes:

a. Problems which are not specific of the KOR activities context but omnipresent in the world today, such as: Digital Segregation, Pornography, Discharge of electronic garbage, Professionals being replaced by technology, Violence;

b. Problems which directly relate to KOR professional activities, such as: Surveillance, Censorship, Lack of cultural warrant, Negligence, Informational Directness, Professional Inefficiency, Misrepresentation, Racism, Ambiguity, Marginalization, Impartiality or Neutrality Belief, Idiosyncrasy, Inaccessibility to information, Biased Terminology, Inadequate Translations.

In this context, a great dispersion of problems was verified and only the problem of Surveillance showed a significant incidence itself (12%).

**Conclusion**

The analysis showed that the prevailing values — and the problems derived from them — partially reveal two complementary dimensions: one concerning the respect to diversity and the other concerning the specificity warrant, which may prove the relevance of the theoretical principles put forward by Hudon (1997), Beghtol (2002, 2005) and Gar-
cía Gutiérrez (2002) who propose a transcultural ethics in information mediation which shall reflect diversity (whose precision may be closely related to the multilinguism issue), making use of tools which can guarantee respect to the cultural dimension of knowledge representation.

All these discussions lead us to reflections on KO education, suggesting that the focus must not only be set on content issues but also on social (and consequently ethical) issues, because the subject representation of information, as an activity which aims at making information accessible to every and all kind of user of a specific information system, should do its job in such a way as to make every user of this system see himself reflected in its services. In case where a violation of this ethical principle is felt by a user, he/she will certainly leave the system, by considering it not viable nor reliable for not having met his/her needs, not respected his/her moral rights and values.

Finally, in the context of subject representation of information, it is necessary to point out the need of further studies investigating violations of ethical values and also analyzing the axiological core of the area once it has been searching for sound foundation of its epistemological core. Furthermore, the need of carrying out further research of this nature becomes more and more evident when we think of issues related to ethical values and ethical problems being present not only in traditional information contexts as libraries, but also in digital information environments such as databases in which the problem of ethics assumes a significantly more complex dimension.

Acknowledgement
This research was supported by the following Brazilian institutions: CNPQ — National Council for Scientific and Technological Development and FAPESP — São Paulo Research Foundation.

References


Olson, H. A. (2002). The power to name: locating the limits or subject representation in libraries. Dordrecht: Kluwer Academic Publisher.


Information Filtering as a Knowledge Organization Process

Techniques and Evaluation

Abstract
In this study, we are concerned with a field which represents an intellectual, social, and economic practice, strongly linked to a semi-automatic knowledge organization. Competitive Intelligence (CI) is characterized by two major distinctive features: transition from the classical activity of Information Retrieval to organised Information Filtering, then conversion of filtered information into Knowledge to help decision making.

In the paper, we first show that information filtering systems may be considered as semi-automatic knowledge organization devices in the business intelligence context. Then, we point out how the technical dimension of the system must be arranged with the user dimension in order to approach a real relevance. Finally, we present the overview of the InFile evaluation campaign which represents an attempt to validate our approach.

1. Introduction
Beyond the various ways of defining and explaining what knowledge is, the Knowledge Economy represents a major concern for the specialists of the domain (e.g., researchers, practitioners, economists). This economy can not grow without paying attention to the various steps of the “knowledge chain”, from automatic or human data acquisition to knowledge organization and its different uses (e.g., documentation, competitive intelligence, knowledge management).

The Knowledge Economy also faces problems of information overload in the digital age (e.g., proliferation of resources and supports, diversification of formats and structures, increase in volumetry and number of users, multilingualism requirement, and the emergence of new editorial practices). Dealing with the consequence of overproduction means to develop and use new technologies such as clustering, push, filtering, cartography and so on with a number of components (e.g., linguistic, statistic, structural). However, the mediation of these technologies is not without complexity and requires a combination of dimensions, not always obvious, between two dimensions of knowledge organization (Kolmayer, 1999). They are:
• the technical dimension which is based on different conceptual models, various technical environment and resources, and
• the user dimension which is closely linked to culture and philosophy, knowledge and know how in the field, practices and individual interests, preferences and subjectivity.

The profitability of systems depends on the compatibility and interactivity between these two dimensions. In our paper, we focus on two case studies: the case of competitive intelligence as a sub-field of knowledge economy and the case of filtering systems as mediation tools for knowledge organization.

We present how filtering devices with their various by-products can be exploited in an activity of competitive intelligence and business intelligence process. Then, we discuss the two different perceptions of relevance, according to the technical approach and the user-oriented approach, in order to find evaluation criteria which combine these two different knowledge organizations. We conclude with a presentation of the evaluation protocol of the InFile campaign devoted to information filtering systems.
2. Filtering devices in a business intelligence process

Because knowledge represents a large part of the intangible goods of each company and a way to compete more efficiently, strategic information systems and knowledge management systems are of great importance. Because of the proliferation of electronic information and the underlying difficulty to manage this information in a relevant way, the usual solution to managing information is to drastically reduce the volume of documents available to the end users using abstracting or filtering process (Chaudiron & Fluhr, 2001).

Filtering systems successfully function by separating relevant and non-relevant documents in an incoming stream of textual information. According to Belkin and Croft (Belkin & Croft, 1992), an information filtering system is a system designed to manage unstructured or semistructured data. We may consider that, nowadays, these systems also manage unstructured data such as pure textual documents. Information filtering systems deal primarily with textual information, and involve large amounts of data incoming through permanent streams such as newswire services. Filtering is based on individual or group information profiles which we assume are used to represent consistent and long-term information needs. From the user point of view, the filtering process is usually meant to extract relevant data from the data streams, according to the user profiles.

Information filtering systems may be used in different business environments: for example, text routing involves sending relevant incoming data to individuals or specific groups, categorization process aims at attaching one or more predefined categories to incoming documents, or anti-spamming tries to remove “junk” e-mails from the incoming e-mails. In the context of competitive intelligence, information filtering may be considered as a very specific subtask of the information management process. In this approach, the information filtering task is very similar to Selective Dissemination of Information (SDI), one of the original and usual functions assumed by documentalists and, more recently, by other information intermediaries such as technological watchers or business intelligence professionals.

As many authors mentioned it, information filtering is a key issue in the business or competitive intelligence process. In the different models of the competitive intelligence cycle, we constantly find the “information acquisition” step as a main task of the whole process. According to Boutheillier and Shearer (Boutheillier & Shearer, 2003), a specific subtask of the “information acquisition” task is to “filter [the content] in order to retain the desired information and discard unwanted information”. Filtering means examining whether the collected information addresses the needs, topics and requirements that were identified previously. For AFNOR (1998), the French official body in charge of the normalization process which provided the de jure standard concerning the watch services in business environment, the whole cycle of competitive intelligence implies 8 steps among which the “information gathering and selecting” task.

While information retrieval systems deal with a relatively stable document set and constantly new queries, in contrast, information filtering (also known as routing or selective dissemination of information), the queries (or profiles) are fixed over time and new documents are constantly added to the initial set. A good example of this situation is a system that filters wires coming from news agencies such as Reuters, Bloomberg or Agence France Presse (AFP).

The filtering task can be carried out by different means, some by automatic tools such as filtering software, but not all. More generally, filtering is a process of organizing information according various criteria. This process may be done by a single person (cognitive
filtering) or in a cooperative way within a group or a community (social filtering). Cognitive filtering is a process that uses the content of information to define the user profile. The profile contains information concerning the user’s interests and supposed information needs. The filtering technique matches incoming documents with the profile, and the global performance of the system is evaluated through feedback from users. Information is recommended on the basis of feedback, recommendations, and the cognitive profiles of ‘similar’ users. In this respect, social filtering is also content-based but this model mainly uses social parameters such as a user’s education, occupation, knowledge and experience as well as preferences and habits. The system also assumes that users with matching social parameters will also share preferences and habits. This relies on the creation of user stereotypes, with sets of rules applied to each stereotype. This kind of systems usually provides ranking filtering so irrelevant items are not discarded but given a low ranking.

The process of filtering may be based on the characteristic of the document such as the words it contains (keywords which may be terms or concepts, named entities), syntactic patterns which represent events (mergers and acquisitions of companies for example) or based on a complete linguistic analysis of the document. Another way to filter documents, commonly referred to as “recommender systems” is to base the filtering process on annotations made to the documents by other users. This distinction between content based and annotation based filtering accords partially with the former distinction between cognitive and social approaches. With the development of the collaborative filtering mechanisms within specialized communities (professional or not), the question of the user model is redefined. User models are usually hand-crafted and/or refined with machine learning techniques using explicit or implicit relevance feedback.

Another way to consider characteristics or types of information filtering is to distinguish between text classification and text clustering. These techniques have been reported extensively in the traditional IR literature. Text classification is the classification of textual documents into predefined categories (supervised process) and text clustering groups documents into categories defined dynamically, based on their similarities (unsupervised process). In classification, categories are first determined (such as the Library of Congress Classification, the Dewey Decimal Classification or the Yahoo! categories) and the incoming information (or documents) are filtered according to an existing structured hierarchy. In clustering, categories are revealed in a bottom-up approach as result of grouping objects based on similarities. Both classification and clustering are filtering techniques.

3. A relevance of filtering based on system knowledge and user knowledge
The information filtering process differs from information retrieval by several aspects but the two processes face the problematic question of relevance concerning the results given by the systems (Belkin & Croft, 1992) (Berti-Equille, 2002).

In the case of information retrieval, the general organization of relevance assessment consists of comparing, in a single session, the query formulated in the search language of the system with the index representing the texts collection. The matching can be exact (boolean model) or optimal (vectorial or probabilistic model) with possibly a weighted answers ranking. Several tests of improvement by techniques of requests reformulation have been proposed (Ben-Ali & Timimi, 1999), but this approach is still faced with problems of adequacy between the expression of the information requirement and the information presentation. In order to reduce these limitations, another technique consists in evaluating the texts returned in this first session by the user, then reinjecting in a second
session, new relevance criteria. This technique often involves a modification of the query and its progressive refinement by a process of “relevance feedback”.

In the case of information filtering, the user formulates what is required (positive profile) and what is not required (negative profile) in a dynamic and regular information flow, using a representation of its relatively stable centers of interest on the long run. Several tests and techniques were implemented to improve the performances of filtering systems (e.g., adaptive filtering based on the progressive and iterative training, passive collaborative filtering based on the analysis of the user’s behaviors, active collaborative filtering based on user comments or analysis). However, the relevance question still remains a big concern.

We may point out two different approaches of relevance (Denos, 1997) related to two different knowledge sources: relevance to a subject and relevance to a user.

The first “system-oriented” is based on the topic adequacy (topicality) between required information and information returned. It remains formal and mechanical and depends on the correspondence made by the system between the presentation of the request and that of the database. The second “user-oriented” is based on the decision of the user to accept or reject the information collected. It remains difficult to identify and considered too ambiguous or multifaceted to be formalised (Brouard & Nie, 2000). The user decision is mainly related to his explicit knowledge that is organized, in a visible way, in the form of profiles, and especially to his tacit knowledge, organized but in an invisible way, in his memory, his practices and his behavior.

The system efficiency not only depends on the topic exactitude of the question-answers (objective answers and modelisable organization) but also on adequacy between responses and user requirement (subjective answers and formalisable difficult process).

According to Saracevic (1975), it is difficult to between these two kinds of relevance which are complementary. Green (1995) considers that the user is the real judge of relevant document but, on the other hand, it is perhaps not the best because he does not necessarily have the knowledge required to evaluate the relevance of the document. Other researchers, Schamber (1991) and Barry (1994) try to determine an inventory of the useful criteria in the evaluation of the relevance by the user1.

In theory, the good way to find out useful criteria to evaluate systems seems to combine the system-oriented relevance criteria and the user-oriented relevance criteria. That is what we presently try to do in the InFile project, taking into account user preferences based on observations of what we call the “ground truth”.

4. Main features of the InFile Evaluation Campaign
The InFile2 evaluation campaign (INformation, FILtering, Evaluation) is a cross-language adaptive filtering evaluation campaign, sponsored by the French National Research Agency. The campaign is organized by the CEA-LIST, ELDA and the University of Lille3-GERiiCO. It has an international scope as it is a pilot track of the CLEF3 2008 campaigns. For those familiar with TREC’s filtering tasks, the InFile campaign is similar to the TREC-11 filtering track with some characteristics (Robertson & Soboroff, 2002). InFile mainly consists of an adaptative filtering task which tries to simulate an on line

---
1. Document content and source, user’s philosophy and preferences, other sources (consensus, external verification), document cost and accessibility.
crosslingual filtering process. English, French and Arabic were concerned by the process but participants could have been evaluated on mono or bilingual runs.

As a consequence of what we have previously said concerning the information filtering process in sections 2 and 3, we paid a particular attention to the context of use of filtering systems by real professional users. Even if InFile is mainly a technological oriented campaign, we constantly tried to adapt the protocol and the metrics, as close as possible, to the so-called “ground truth”. In respect with that, the global features of InFile are:

**Corpora**

A newswires corpus was provided by the Agence France Presse (AFP). This is a collection of about 1.4 million newswires (10 GB) selected from a 3-year period. Newswires are available in the three mentioned languages but are not translations from a language to another.

We prepared a set of 50 profiles covering two different categories: the first group deals with general news and events concerning national and international affairs, sports, politics... and the second one deal with scientific and technological subjects. In order to be as close as possible to the “ground truth”, profiles were constructed by competitive intelligence professionals from INIST⁴ (the French Institute for Scientific and Technical Information Center), ARIST Nord Pas-de-Calais⁵ (Agence Régionale d’Information Stratégique et Technologique), Digiport⁶ and ONERA⁷ 30 of these are general profiles and 20 are scientific profiles. The practitioners constructed both the English and the French versions of the profiles while the Arabic version was translated by native speakers.

**Relevance judgments**

The relevant set of documents was constructed in two phases, a pre-submission phase and a post-submission phase of judgements. Extensive searches using different retrieval systems were conducted at ELDA after the elaboration of the profiles. In this pre-submission phase, both the professional involved in the definition of the profiles and other assessors made relevance judgments on the outputs of the systems. This process included several feedback stages. After one round of such assessment, relevance information was used to improve the profiles and another round of assessment was made. In a post-submission phase, additional relevance judgments are planned to be made by the assessors after submission of results by the participants, on the documents taken from the pooled submissions for each profile. It will allow us to identify additional relevant documents that could have been not found by the assessors at the previous stage.

**Protocol and metrics**

In order to minimize a human intervention during the test, the evaluation task was performed using an automatic interrogation of participating systems with a simulated user feedback but systems were allowed to use the feedback at any time to increase performance. For each profile, systems were given a Boolean decision for each document. Due to the many possible runs, participants were also asked to fill out a form to specify which languages and which kind of profiles they wanted to be evaluated on.

---

Three different metrics have been retained:

- *Progression measure* (or evolutivity) which measures the ability of a system to improve itself from the relevance feedback;
- *Originality measure* which measures the fact that a system is the only one to retrieve some relevant documents;
- *Anticipation measure* which measures the ability of the systems to retrieve the first relevant document; this measure is very close to real conditions of use when it is important to extract “low signals” from an incoming flow of information.

These metrics try to take into account user information behavior during the relevance judgment phase. The metrics have been elaborated after discussions with CI practitioners. They surely do not fit exactly with the real conditions of use, but they can be considered as a first attempt to match with these conditions.

5. Conclusion

At this time, the real test of the InFile campaign did not start yet, so we are not able to present the results of the comparative evaluation of the participants, but the first goal has been achieved. This goal was to define an evaluation protocol paying attention to a real context of use.

Information filtering systems can be considered as a case study to demonstrate how it is possible to deal with a user evaluation referring to cognitive and psychosocial influences and a technical-functional assessment in a unified approach, in order to evaluate systems.

References


Retrieving Terminological Information on the Net  
Are Linguistic Tools Still Useful?

Abstract
In this paper we present a comparative evaluation of the effectiveness of search engines (SE) and linguistic tools (LT) to retrieve terminological information from the net, in the context of specialized translation tasks. For achieving this goal, an experiment with translators has been carried out. The results indicate that SE are more effective than LT in situations where the answer is partially ignored by the translator (i.e. the translator is hypothesizing one or several possible answers in the target language before searching). On the other hand, LT have not been either more appropriate in situations where the translator showed total ignorance of the possible answers before searching.

1. Introduction
In this paper we want to verify the qualities of search engines (SE) as effective tools for solving terminological problems in contexts of specialized translation. *A priori*, on the basis of our experience, we suggest that SE might be more effective than LT conceived and used by professional translators, such as terminological databanks and bilingual dictionaries. With the purpose of validating this hypothesis we classify the information needs of specialized translators based on their previous knowledge of the possible answers and therefore their expectations. Next we carry out an experiment with translators. This allows us to determine the effectiveness of both types of tools for different types of queries.

2. The concept of “information needs” in the domain of translation
A literature review in this domain led us to the conclusion that there are no typologies of information needs established for the purposes of specialized translation. Thus, on the basis of our previous experience analyzing translation problems as translation teachers, we considered two types of situations in which the translator browses the Web to find an equivalent term that he does not know for sure (or he doesn’t know at all). According to this, questions are divided into two main groups:

   A. Questions made from a total ignorance of the answer: the translator does not know the equivalence or the possible equivalences of the term he is searching for. This group of questions is labelled “absolute lack of knowledge” (ALK).

   B. Questions made from a partial ignorance of the answer, i.e., an intuition of the existence of one or more equivalences in the target language, which have to be checked in real usage. These questions will be labelled “partial lack of knowledge” (PLK).

In the first case, the logical point to start with will be linguistic tools (LT) that contain equivalences in source and target language, stored in a lexical database. In the second case, the starting point is more likely accessing real texts in the target language that can validate the intuitions about the answers. The Web is used here as a huge textual corpus, and SE are the means of browsing the corpus.

3. Online tools for translation
Terminological databases and databanks (TDB) have been, from the early seventies on (Eurodicautom, 1973), one of the most important and reliable sources of information for
many professional groups dealing with specialized lexical information (translators, technical writers, lexicographers, information scientists). TDBs aim at recording specific meanings associated with different subject fields, precise definitions, controlled denominative variation (synonyms, abbreviations, alternate spellings, systematic or scientific names, symbols) and equivalents in other languages. In addition, other kinds of sociolinguistic and sociocultural data are also recorded, as geolctal variation, usage notes or conceptual information related to a more restricted language community. All this information is organized and structured into a lexical database, under an onomasiological orientation, i.e. focusing on the concept: every record of a TDB contains all the information related to that concept (definition, contexts, alternatives, equivalents in other languages). This rigid structure in records and fields is supposed to provide a safe and reliable access to the information sought.

On the other hand, the proliferation of electronic dictionaries on the Web has caused the appearance of a mixed type of linguistic resource, which might be situated between a lexicographical resource (bilingual dictionary) and a specialized forum for professionals. In these sites, registered users, mainly professional translators, technical writers and language teachers, complement the information contained in these base dictionaries with contributions dealing with specialized terms, equivalences and other kinds of linguistic and grammatical questions. As opposed to the mentioned LT, there is an increasing use of SE as they are seen as the door to a wide variety of resources.

Table 1 shows a comparison between main features of both resources.

<table>
<thead>
<tr>
<th>SE</th>
<th>LT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Updated</td>
<td>Constant</td>
</tr>
<tr>
<td>Corpus</td>
<td>Not filtered, all the pages indexed by the SE, including some LT accessible as plain text</td>
</tr>
<tr>
<td>Linguistic information</td>
<td>Collocation, Frequencies of usage, Multilingual, Multicultural</td>
</tr>
<tr>
<td></td>
<td>Multidomain</td>
</tr>
<tr>
<td>Reliability</td>
<td>Not controlled, but statistically relevant</td>
</tr>
</tbody>
</table>

### 4. Experiment

For proving the hypothesis presented above and to fulfill the objective of this investigation, we have completed an experiment that was carried out during January and February 2008. The experiment was designed from a set of information needs for specialized translation according to two categories — “absolute lack of knowledge” questions and “partial lack of knowledge” questions.

### 4.1. User’s profile

The experiment was carried out by a group of 16 students at the Faculty of Translation and Interpretation, Universitat Pompeu Fabra, Barcelona. All the students had a strong background in Scientific and Technical Translation, were all at 4th year, their native languages being Spanish or Catalan. Before we completed the test with the sample of 16 users, 2 users of advanced profile made all the tasks without time constraints, and gave us feedback about possible misunderstandings, which allowed us to make some adjustments in the forms and the questions in order to improve the reliability of the test.
4.2. Selecting SE and LT
Two SE and two LT were selected. SE were Google and Yahoo, as they currently are the most generalist SE used anywhere. The selected LT were IATE and Word Reference. Both resources were created for storing lexical equivalences between languages and both are also very popular resources among translators. The first one is a terminological data bank and the second one is a dictionary. All resources are freely accessible on-line, multilingual and multidomain. For the searches with the SE, the users of the experiment were allowed to use all the strategies they knew for refining each search: — Limiting the search to pages from Spain — Limiting the search to pages written in Spanish — Indicating the word “translation” next to the searched term — Indicating the word “glossary” next to the searched term.

4.3. Selecting texts
Two scientific texts in English were selected:

**Figure 1. Scientific texts selected for the experiment**


The first text it is a specialized text that contains recommendations from an international agency in the domain of health care, and is addressed to physicians willing to work in the field, in surveillance and monitoring tasks. The second text is a review article, published in an international journal that presents a discussion about certain concepts in the domain of cellular biology. Both texts are written by specialists and are addressed to specialists, and we consider them representative of the scientific discourse in English language.

4.4. Previous questions to the tasks
After reading the texts, the users had to propose a translation (one or several equivalent terms) for the words and expressions that were requested. Besides, the users were also requested to indicate the level of success they think they might achieve for each of the questions, in order to evaluate their previous knowledge about the problems raised, and their expectations before carrying out the search.

**Figure 2. Example of previous question that had to be answered**

<table>
<thead>
<tr>
<th>Tarea G1T1PA1</th>
</tr>
</thead>
<tbody>
<tr>
<td>¿Cuál es la equivalencia en castellano de <em>finger stick</em>?</td>
</tr>
<tr>
<td>Respuesta: _______________ Nivel de acierto esperado (0–10): ______</td>
</tr>
</tbody>
</table>

4.5. Tasks
Each user was asked to give an answer to a set of translation problems selected from the texts presented above. The time to answer each question was limited to 10 minutes. Should they have no answer for the problem raised, they had to skip to the following question. The selected questions, organized according to the categories above mentioned, included the most common problems occurred in translating scientific texts.
• Questions arose from an absolute lack of knowledge: the translator does not know the equivalence or the possible equivalences. Example: *Which is the equivalent of 'serosurveillance'?*

• Questions arose from a partial lack of knowledge: the translator is capable of hypothesizing one or more equivalences in the target language, although these options need to be checked. Example: *For mechanistically, it must be said “mecánicamente, mecanicistamente or mecanicisticamente”*?

The tasks were distributed to users in two groups (G1, G2). Each user had to work with both texts (T1, T2) according to the following table:

<table>
<thead>
<tr>
<th>USERS GROUP 1</th>
<th>USERS GROUP 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Text 1</strong></td>
<td></td>
</tr>
<tr>
<td>1º LT</td>
<td>1º SE</td>
</tr>
<tr>
<td>2º SE</td>
<td>2º LT</td>
</tr>
<tr>
<td><strong>Text 2</strong></td>
<td></td>
</tr>
<tr>
<td>1º SE</td>
<td>1º LT</td>
</tr>
<tr>
<td>2º LT</td>
<td>2º SE</td>
</tr>
</tbody>
</table>

### 4.6. Answers to the tasks

For each of the eight questions to be answered, users were requested to look for the solution in each resource and evaluate the satisfaction of the information sought. This is an example of the questionnaire:

**Figure 3. Questionnaire**

<table>
<thead>
<tr>
<th>Tarea G1T1PA1</th>
<th>¿Cuál es la equivalencia en castellano de <em>finger stick</em>?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IATE</strong></td>
<td>¿Ha obtenido la respuesta? Sí / En parte / No Satisfacción (0–10):</td>
</tr>
<tr>
<td><strong>Yahoo!</strong></td>
<td>¿Ha obtenido la respuesta? Sí / En parte / No Satisfacción (0–10):</td>
</tr>
<tr>
<td><strong>Respuesta:</strong></td>
<td>Estrategia de búsqueda que le ha dado la respuesta:</td>
</tr>
<tr>
<td><strong>Posición del sitio web en el ranking de resultados (1–10):</strong></td>
<td></td>
</tr>
<tr>
<td><strong>URL:</strong></td>
<td>http://</td>
</tr>
<tr>
<td><strong>Comentarios:</strong></td>
<td></td>
</tr>
</tbody>
</table>

In the searches carried out in SE, users had to indicate in addition: in which position was the answer ranked, the searching strategy, and the URL of the Web site where the answer was obtained.

### 5. Results

The experiment has given us data that have to be carefully analyzed. For the purposes of the current research, we have first analyzed the variable of effectiveness. By effectiveness we mean the capacity a tool has to provide the needed of information.

In order to determine the degree of effectiveness of the tools we wanted to compare, we have used the answer to the question “has obtained the answer” and we have entered the answers “yes”, “partly” and “no” so that each affirmative answer has supposed a point, each answer “partly” has been valued half a point and each negative answer has not added any point. The scores obtained by type of question and type of tool are presented in figure 4.
Figure 4. Values obtained in effectiveness for the linguistic tools (LT) and SE to questions for which the user didn’t know the answer totally (ALK) or partially (PLK)

According to these results, SE are more effective than LT for both types of questions, no matter if the user showed ALK or PLK, but are more effective in the questions of the second type (PLK), as we raised in the initial hypothesis. Presenting these data in percentage, we can see how SE have a higher percentage of positive answers (table 3).

Table 3. Percentage of answers obtained for each type of tool in each type of question

<table>
<thead>
<tr>
<th></th>
<th>ALK</th>
<th>PLK</th>
</tr>
</thead>
<tbody>
<tr>
<td>LT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>25%</td>
<td>28%</td>
</tr>
<tr>
<td>Partly</td>
<td>31%</td>
<td>28%</td>
</tr>
<tr>
<td>No</td>
<td>44%</td>
<td>44%</td>
</tr>
</tbody>
</table>

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>63%</td>
<td>72%</td>
</tr>
<tr>
<td>Partly</td>
<td>9%</td>
<td>16%</td>
</tr>
<tr>
<td>No</td>
<td>28%</td>
<td>13%</td>
</tr>
</tbody>
</table>

Figure 5. Percentage according to the different usages of SE tools
Looking at these results where the SE seem to win both types of questions with enough
difference with respect to LT, we introduced a second variable that would still modify the
values presented.

For the purposes of specialized translations, SE can be used either as a textual corpus,
as a corpus to see relative frequencies of usage, but also as pointers to other LT resources.
We decided to review in how many cases the answers obtained in SE brought the user to a
linguistic resource (i.e., a dictionary or a terminological data bank). This happened to be
the 16% in the case of ALK questions and 0% in PLK questions (figure 5).

If we increase the values presented in figure 4 considering that there has been a 16%
percentage of positive answers in SE where the resource pointed to a LT, we can see the
difference in the following values (figure 6). After the correction, the score of the answers
for which the user has a partial ignorance has increased in the evaluation of the LT, and
has decreased in the evaluation of the effectiveness of SE. As for the answers for which
the user had a partial ignorance, there has been no variation, thus no correction.

If in figure 4 we could affirm that SE presented higher degree of effectiveness than the
LT to solve translation problems, and this degree is slightly higher for the questions in
which the user has a partial knowledge of the answer (PLK), the correction considering
that in some occasions the SE points to a LT show that SE are better valued in PLK ques-
tions (8/10 points, as opposed to 4/10 points of LT), whereas for ALK questions both
types of tools maintain similar lower values.

Therefore, before and after the correction the hypothesis of this research work has been
validated. SE are more effective to solve translation problems in which the ignorance of
the answer is partial. However, it has not been possible to verify that LT would be more
appropriate in the cases where the ignorance of the answer is total, because the values of
these answers before correction are much lower than the values obtained by SE, and after
the correction the values are practically equal.
6. Future work
The results obtained allow us to further investigate considering qualitative variables we have in the questionnaires:

• Previous knowledge of the answers. Being aware of the previous user’s knowledge for each one of the answers is a relevant point because it will indicate if the questions for which we estimate “total ignorance” deserve such label. There is no doubt in the case of the questions labelled as “partial ignorance” since in the question we gave several possible answers.

• Satisfaction. This value will be calculated on the basis of the evaluation given by the users for each answer. We will be able to obtain a score (from 0 to 10) for both types of questions (ALK and PLK) and type of tools (LT and SE).

As far as the methodology is concerned, in a further stage of the research we will have a user group making the tasks with direct observation from our part. Think aloud techniques will be used, and sessions will be recorded to be able to study carefully the steps taken in the search process, having the opportunity of revising comments made by the user, and thus determining precisely the time elapsed and the number of clicks made for answering each question.
List of Contributors

Abbas, June
Department of Library and Information Studies,
Graduate School of Education,
State University of New York at Buffalo
547 Baldy Hall
Amherst, NY 14260 USA
abbasjm@buffalo.edu

Abreu, Amelia
University of Washington
MGH 370, Box 352840
Seattle, WA 98195-2840 USA
aabreu@u.washington.edu

Alberts, Inge
École de bibliothéconomie et des sciences de l’information,
Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal, QC H3C 3J7 Canada
inge.alberts@umontreal.ca

Andersen, Jack
Royal School of Library and Information Science
Birketinget 6
København S DK-2300 Denmark
jan@db.dk

Arsenault, Clément
École de bibliothéconomie et des sciences de l’information,
Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal, QC H3C 3J7 Canada
clement.arsenault@umontreal.ca

Bertrand-Gastaldy, Suzanne
École de bibliothéconomie et des sciences de l’information,
Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal, QC H3C 3J7 Canada
suzanne.bertrand-gastaldy@umontreal.ca

Bosch, Mela
Journalism Department,
La Plata National University
La Plata, Buenos Aires Province Argentina
melabosch@europe.com

Broughton, Vanda
School of Library Archive & Information Studies,
University College London
Gower Street
London WC1E 6BT UK
v.broughton@ucl.ac.uk

Buckland, Michael
School of Information,
University of California,
Berkeley, CA 94720-4600 USA
buckland@ischool.berkeley.edu

Chambers, Thane
School of Library and Information Studies,
University of Alberta
3-09 Rutherford South
Edmonton, AL T6G 2J4 Canada
thane.chambers@ualberta.ca

Campbell, D. Grant
Faculty of Information and Media Studies,
University of Western Ontario
London, ON N6A 5B7 Canada
gcampbel@uwo.ca

Chan, Lois Mai
School of Library & Information Science,
University of Kentucky
320 Little Fine Arts Library
Lexington, KY 40506-0224 USA
loischan@uky.edu

Chaudhry, Abdus Sattar
Division of Information Studies,
Wee Kim Wee School of Communication & Information
Technological University
Nanyang, Singapore
aschaudhry@ntu.edu.sg

Chaudiron, Stéphane
Laboratoire GERIICO,
Université Charles de Gaulle-Lille3
Domaine universitaire du « Pont de Bois »,
rue du Barreau — BP 60149
Villeneuve d’Ascq
Cedex 59653 France
stephane.chaudiron@univ-lille3.fr

Chen, Chaomei
College of Information Science & Technology,
Drexel University
3141 Chestnut Street
Philadelphia, PA 19104 USA
Chaomei.Chen@ischool.drexel.edu

Chen, Xu
Institute for Library and Information Science
Humboldt University in Berlin
D-10099 Berlin, Germany
chenxu@rz.hu-berlin.de
Chuttur, M. Yasser  
School of Library & Information Science,  
Indiana University Bloomington  
011 Wells Library, 1320 E. 10th St.  
Bloomington, IN 47405 USA  
mchuttur@indiana.edu

Corrado, Edward M.  
Binghamton University Libraries  
P.O. Box 6012, Vestal Parkway East  
Binghamton, NY 13902-6012 USA  
eccorardo@ecorrado.us

Dahlstrom, Mats  
Swedish School of Library and Information Science,  
UC Borås / Göteborg University  
Allégatan 1  
Borås SE-50190 Sweden  
mats.dahlstrom@hb.se

DiMarco, John  
Division of Mass Communication,  
St. John’s University  
8000 Utopia Parkway  
Queens, NY 11439 USA  
dimarcoj@stjohns.edu

Doua, Thomas M.  
Graduate School of Library and Information Science,  
University of Illinois, Urbana-Champaign  
501 E. Daniel Street, MC-493  
Champaign, IL 61820-6211 USA  
tdoua2@uiuc.edu

Feinberg, Melanie  
University of Washington,  
Roosevelt Commons Building  
4311 11th Ave NE, Box 354985  
Seattle, WA 98195-4608 USA  
feinberg@u.washington.edu

Fels, Deborah  
Ted Rogers School of Information Technology Management,  
Ryerson University  
350 Victoria Street  
Toronto, ON M5B 2K3 Canada  
dfels@ryerson.ca

Fernández-Molina, Juan Carlos  
Facultad de Comunicación y Documentación.  
Universidad de Granada  
Granada 18071 Spain  
jcfernandez@ugr.es

Furner, Jonathan  
Graduate School of Education & Information Studies, UCLA  
300 Young Dr N, Mailbox 951520  
Los Angeles, CA 90095-1520 USA  
furner@gseis.ucla.edu

George, Nicolas  
School of Library & Information Science,  
Indiana University Bloomington  
011 Wells Library, 1320 E. 10th St.  
Bloomington, IN 47405 USA  
nlgeorge@indiana.edu

Gnoli, Claudio  
Dipartimento di Matematica,  
University of Pavia  
Via Ferrata 1  
Pavia I-27100 Italy  
 gnoli@aib.it

Goodrum, Abby A.  
Faculty of Communication and Design,  
Ryerson University  
350 Victoria Street  
Toronto, ON M5B 2K3 Canada  
agoodrum@ryerson.ca

Green, Rebecca  
Dewey Decimal Classification,  
OCLC Online Computer Library Center, Inc.  
Library of Congress,  
Decimal Classification Division,  
101 Independence Ave., S.E.  
Washington, DC 20540-4330 USA  
greenre@oclc.org

Guimarães, José Augusto Chaves  
Departamento de Ciência da Informação,  
São Paulo State University  
Av. Hygino Muzzi Filho, 737  
Marilia, SP 17525-900 Brazil  
guimajar@marilia.unesp.br

Guo, Lijiang  
School of Library & Information Science,  
Indiana University Bloomington  
011 Wells Library, 1320 E. 10th St.  
Bloomington, IN 47405 USA  
lijguo@indiana.edu

Hajdu Barat, Ágnes  
Head of Library and Information Science Department  
University of Szeged Gyula Juhász  
Faculty of Education (SZTE JGYPK)  
Szeged, Szilléri sgt. 12. 6723 Hungary  
hajdu@jgypk.u-szeged.hu

Hajibayova, Lala  
School of Library & Information Science,  
Indiana University Bloomington  
011 Wells Library, 1320 E. 10th St.  
Bloomington, IN 47405 USA  
lahajibay@indiana.edu
Hansson, Joacim
Swedish School of Library and Information Science,
UC Borås / Göteborg University
Allegatan 1
Borås SE-50190 Sweden
joacim.hansson@hb.se

Hibbard, Ellen
Center for Learning Technologies,
Ryerson University
350 Victoria Street
Toronto, ON M5B 2K3 Canada
ehibbard@ryerson.ca

Hill, Amanda
University of Manchester
59 Carnrike Road
Consecon, ON K0K 1T0 Canada
amanda.hill@manchester.ac.uk

Hjørland, Birger
Royal School of Library and Information Science
Birketinget 6
København S DK-2300 Denmark
bh@db.dk

Howarth, Lynne C.
Faculty of Information Studies,
University of Toronto
140 St George Street
Toronto, ON M5S 3G6 Canada
lynne.howarth@utoronto.ca

Hudon, Michèle
École de bibliothéconomie et des sciences de l’information,
Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal, QC H3C 3J7 Canada
michele.hudon@umontreal.ca

Ibekwe-SanJuan, Fidelia
College of Information Science & Technology,
Drexel University
3141 Chestnut Street
Philadelphia, PA 19104 USA
ibekwe@drexel.edu

Jacob, Elin K.
School of Library and Information Science,
Indiana University Bloomington
011 Wells Library, 1320 E. 10th St.
Bloomington, IN 47405 USA
ejacob@indiana.edu

Khoo, Christopher S. G.
Division of Information Studies,
Wee Kim Wee School of Communication & Information Technologies
National University of Singapore
assgkhoo@ntu.edu.sg

Kipp, Margaret E.I.
Palmer School of Library and Information Science,
Long Island University
720 Northern Boulevard
Brookville, NY 11548-1300 USA
margaret.kipp@gmail.com

Kjellman, Ulrika
Library and Information Science,
Department of ALM, Uppsala University
Box 625
Uppsala 751 26 Sweden
Ulrika.Kjellman@abm.uu.se

La Barre, Kathryn
Graduate School of Library & Information Science,
University of Illinois at Urbana-Champaign
501 E. Daniel Street
Champaign IL 61820-6211 USA
klabarre@uiuc.edu

Lee, Hur-Li
School of Information Studies,
University of Wisconsin-Milwaukee
PO Box 413
Milwaukee, WI 53201 USA
hurli@uwm.edu

Lima, Gercina Angela Borém Oliveira
School of Information Sciences of UFMG
Av. Exp. Benvindo Belém de Lima, 1055 São Luiz, Belo Horizonte
Minas Gerais 31310-040 Brazil
glima@eci.ufmg.br

Loehrlein, Aaron
School of Library & Information Science,
Indiana University Bloomington
011 Wells Library, 1320 E. 10th St.
Bloomington, IN 47405 USA
aloehrle@indiana.edu

López-Huertas, María J.
Facultad de Comunicación y Documentación.
Universidad de Granada.
Colegio Máximo de Cartuja
Granada 18071 Spain
mjlopez@ugr.es
Marcos, Mari-Carmen
Universitat Pompeu Fabra
La Rambla, 30-32
Barcelona 08002 Spain
mcarmen.marcos@upf.edu

Marcoux, Yves
GRDS, Ecole de bibliothéconomie et des sciences de l’information,
Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal, QC H3C 3J7 Canada
yves.marcoux@umontreal.ca

Mas, Sabine
Laboratoire Tech-CICO,
Université de Technologie de Troyes
12 rue Marie Curie
B.P. 2060, 10010 Troyes Cedex France
sabine.mas@utt.fr

Mazzocchi, Fulvio
Institute for Atmospheric Pollution
National Research Council
Via Salaria Km 29,300 - C.P. 10
00016 Monterotondo staz. (RM) - Italy
mazzocchi@iaa.cnr.it

McIlwaine, Ia C.
School of Library Archive & Information Studies,
University College London
Gower Street
London WC1E 6BT UK
i.mcilwaine@ucl.ac.uk

Ménard, Elaine
École de bibliothéconomie et des sciences de l’information,
Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal QC H3C 3J7 Canada
elaine.menard@umontreal.ca

Milani, Suellen Oliveira
Departamento de Ciência da Informação,
São Paulo State University
Av. Hygino Muzzi Filho, 737
Marília, SP 17525-900 Brazil
suellenmilani@yahoo.com.br

Mitchell, Joan S.
OCLC Online Computer Library Center, Inc.
6565 Kilgour Place
Dublin, OH 43017-3395 USA
mitchelj@oclc.org

Moulaison, Heather L.
Rutgers, The State University of New Jersey,
School of Communication, Information, and Library Studies
4 Huntington Rd.
New Brunswick, NJ 08901 USA
heather.moulaison@rutgers.edu

Mustafa El Hadi, Widad
UFR IDIST Laboratoire GERIICO,
Université Charles de Gaulle-Lille3
Domaine universitaire du « Pont de Bois »,
rue du Barreau — BP 60149
Villeneuve d’Ascq
Cedex 59653 France
widad.mustafa@univ-lille3.fr

Naumis, Catalina
Centro Universitario de Investigaciones Bibliotecológicas,
Universidad Nacional Autónoma de México
Mexico City Mexico
naumis@servidor.unam.mx

Neelameghan, A.
216, 4th Main Road, 16th Cross Road,
Malleswaram West,
Bangalore 560055 India
anm2002@vsnl.net

Ohly, H. Peter
GESIS, Social Science Information Center
Lennestr. 30
Bonn 53113 Germany
peter.ohly@gesis.org

Oliveira Lima, João Alberto de
Universidade de Brasilia, Senado Federal
Brasilia, Senado Federal Brazil
joaoli13@gmail.com

Oliveira, Maria Odaisa Espinheiro de Oliveira
Faculdade de Biblioteconomia,
Universidade Federal do Pará
Rua Augusto Corrêa,
s/n 66075-900 — Belém — Pará — Brazil
odaisa@ufpa.br

Olson, Hope A.
School of Information Studies,
University of Wisconsin-Milwaukee
510G Bolton Hall
Milwaukee, WI 53201 USA
holson@uwm.edu

Ou, Shiyan
Research Group in Computational Linguistics
University of Wolverhampton
Stafford Road
Wolverhampton, UK WV1 1SB
shiyan.ou@wlv.ac.uk

Palmirani, Monica
Università degli Studi di Bologna
Via Zamboni, 33
Bologna 40126 Italy
palmirani@cirifid.unibo.it
Pinho, Fabio Assis
Departamento de Ciência da Informação, São Paulo State University
Av. Hygino Muzzi Filho, 737
Marília, SP 17525-900 Brazil
fabio@fcav.unesp.br

Pinho, Roberto
Universidade de São Paulo
São Paulo SP Brazil
robertopinho@yahoo.com.br

Pisanski, Jan
National and University Library, Slovenia
Turjaska 1
Ljubljana 1000 Slovenia
jan.pisanski@nuk.uni-lj.si

Qin, Jian
School of Information Studies, Syracuse University
Syraucse, NY 13244 USA
jxin@syr.edu

Raghavan, K. S.
DRTC Indian Statistical Institute
Bangalore 560059 India
raghavan@isibang.ac.in

Riley, Jenn
Indiana University Digital Library Program
1320 E. 10th St., W501
Bloomington, IN 47405 USA
jenhrle@indiana.edu

Rizkallah, Élias
GRDS, Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal, QC H3C 3J7 Canada
elias.rizkallah@gmail.com

Rype, Ingebjörg
National Library of Norway
P.o. Box 2674 Solli
Oslo NO-0203 Norway
Ingebjorg.Rype@nb.no

Shaw, Ryan
School of Information, University of California, Berkeley, CA 94720-4600 USA
ryanshaw@ISchool.Berkeley.EDU

Shiri, Ali
School of Library and Information Studies, University of Alberta
3-09 Rutherford South
Edmonton, AL T6G 2J4 Canada
ashiri@ualberta.ca

Simon, Judith
Research Group Philosophy of Science: Cultures and Technologies of Knowledge, Department of Philosophy, University of Vienna
Universitaetsstr. 7
Vienna A-1010 Austria
judith.simon@univie.ac.at

Smiraglia, Richard P.
Palmer School of Library and Information Science, Long Island University, 720 Northern Blvd
Brookville, NY 11548 USA
richard.smiraglia@liu.edu

Spiteri, Louise
School of Information Management, Dalhouse University
Kenneth C. Rowe Management Building, 6100, University Avenue, Halifax, NS B3H 3J5 Canada
Louise.Spiteri@dal.ca

Svanberg, Magdalena
National Library of Sweden
Box 5039
Stockholm SE-102 41 Sweden
magdalena.svanberg@kb.se

Szostak, Rick
Department of Economics, University of Alberta
Edmonton, AL T6G 2H4 Canada
Rick.szostak@ualberta.ca

Tebé, Carles
Universitat Pompeu Fabra
La Rambla, 30-32
Barcelona 08002 Spain
carles.tebe@upf.edu

Tennis, Joseph T.
The Information School, University of Washington
Box 352840, Mary Gates Hall, Ste 370
Seattle, WA 98195-2840 USA
jtennis@u.washington.edu

Timimi, Ismail
Laboratoire GERIICO, Université Charles de Gaulle-Lille3
Domaine universitaire du « Pont de Bois », rue du Barreau — BP 60149
Villeneuve d’Ascq
Cedex 59653 France
ismail.timimi@univ-lille3.fr
Turner, James  
École de bibliothéconomie et des sciences de l’information,  
Université de Montréal  
C.P. 6128, succ. Centre-Ville  
Montréal, QC H3C 3J7 Canada  
james.turner@umontreal.ca

Vargas, Verónica E.  
Mexican Institute of Water Technology  
Paseo Cuauhnahuac 8532  
62550 Jiutepec, Mor. Mexico  
veronicavarsua@prodigy.net.mx

Vitali, Fabio  
Università degli Studi di Bologna  
Via Zamboni, 33  
Bologna 40126 Italy  
vitali@cs.unibo.it

Williamson, Nancy J.  
Faculty of Information Studies,  
University of Toronto  
140 St George Street  
Toronto ON M5S 3G6 Canada  
nancy.williamson@utoronto.ca

Woodcock, Kathryn  
School of Occupational and Public Health,  
Ryerson University  
350 Victoria Street  
Toronto, ON M5B 2K3 Canada  
kwoodcoc@ryerson.ca

Yi, Kwan  
School of Library & Information Science,  
University of Kentucky  
320 Little Fine Arts Library  
Lexington, KY 40506-0224 USA  
kwan.yi@uky.edu

Zacklad, Manuel  
Laboratoire Tech-CICO,  
Université de Technologie de Troyes  
12 rue Marie Curie  
B.P. 2060, 10010 Troyes Cedex France  
manuel.zacklad@utt.fr

Zaher, L’Hédi  
Laboratoire Tech-CICO,  
Université de Technologie de Troyes  
12 rue Marie Curie  
B.P. 2060, 10010 Troyes Cedex France  
zaher@utt.fr

Zhang, James  
College of Information Science & Technology,  
Drexel University  
3141 Chestnut Street  
Philadelphia, PA 19104 USA  
jz85@drexel.edu

Žumer, Maja  
University of Ljubljana  
Faculty of Arts  
Askerceva 2  
Ljubljana 1000 Slovenia  
maja.zumer@ff.uni-lj.si
Subject Index

Numbers
4W (What, Where, When, Who) vocabulary mapping 151–155

A
Administrative electronic document organization 137–142
Alzheimer’s Disease International (ADI) 326
Amazonian Paraense culture 84–89
American Sign Language (ASL) 282, 284–286. See also Keysigns
Archival description, of Web documents 118–123
Astronomical research 307–312
Authorial voice (Concept) 22–27
Automated clustering of concepts 334–336. See also Category clusters, Clustering of concepts, Human clustering of concepts
– Comparison with human clustering 336–338

B
Bias in knowledge organization 256–260
Bibliographical classification, Comparison with scientific classification 235–236, 240
Bilingual thesaurus 70–76. See also Compendex Thesaurus, INSPEC Thesaurus, Thesaurus, Thesaurus construction, Thesaurus development
– Conceptual relationships in 71–76
Bliss Bibliographic Classification 2nd edition (BC2) 43–49
Bloom’s Taxonomy, As learning objects classification structure 15–21

C
Canada, E-mail management practices in public administrations 347–352
Categorization process, Impact on memory 326–330
Category clusters, Use by people with early stage dementia 326–330. See also Clustering of concepts, Human clustering of concepts
Children’s resources, Use of controlled vocabulary 354–359
Chinese bibliographic classification 275–280
Chinese library catalogue (Seven Epitomes) 275–280
CiteULike 320, 322–323
Classical Tamil studies 70–76
Clustering of concepts 333–339. See also Automated clustering of concepts, Category clusters, Human clustering of concepts
Code4Lib community, Use of social tagging in 313–316
Cognition, and everyday classification 249–255
Communities of practice, Use of social tagging in 313–318
Compendex Thesaurus 184–189. See also Bilingual thesaurus, INSPEC Thesaurus, Thesaurus, Thesaurus construction, Thesaurus development
Competitive intelligence (CI), Use of filtering systems in 367–372
Complex concepts, Interpretation of 295–301
Conceptual coherence 3–7
Conceptual models 3–7
– Interoperability of 124–129
– Knowledge-based theories 5–6
– Similarity-based theories 4–5
Controlled semantics 229–233
Controlled vocabularies. See also Uncontrolled vocabularies
– Application of intertextual semantics on 36–41
– Use for children’s resources 354–359
– Use for image indexing 77–82
Corporate websites structure 157–160
Cross-language information retrieval systems 91–96
Cultural identity 1
– In knowledge representation systems 84–89
Cultural markers 56–61
Cultural websites 56–61
Culture, Impact on subject field development 340–345. See also under Databases

D
Darfur crisis, Discourse on 304–305
Databases, As a cultural form 270–271
del.icio.us 314, 316–318
– Tagging patterns in 164–169
Derrida’s theory on logocentrism 226–227
Descriptive practices of archives, For the Web 118–123
Dewey Decimal Classification (DDC). See also under Facets
– Hidden relationships in 8–14
– Relative Index (RI) 8–14
– Use of mixed translations in 98–103
Dewey’s Desimalklassifikasjon 99–103
Digitization practices within national libraries 111–116
Discourse analysis 302–306
Discourse community 302–306
– Linguistic analysis on 304–306

E
E-mail management practices in Canadian public administrations 347–352
E-mail Pragmatic Framework model 347, 349–350
Epistemology 216–221
Ethics in knowledge organization 361–365
– Ethical problems 364
– Ethical values 363–364
Everyday classification process, Role of cognition in 249–255

F
Faceted information infrastructures for websites 105–110. See also Faceted knowledge system, Faceted terminology, Facets
Faceted knowledge system for administrative document organization 137–142. See also Faceted information infrastructures, Faceted terminology, Facets
Faceted terminology, Semi-automatic management of 43–49. See also Faceted information infrastructures, Faceted knowledge system, Facets
Facets. See also Faceted information infrastructures, Faceted knowledge system, Faceted terminology
– Definition 106
– Use in libraries websites 108–109
– Use in museums websites 108–109
– Used in the Universal Decimal Classification (UDC) 50–54
Feinberg’s theory on subjectivity in classification 256–260
Filtering systems in competitive intelligence (CI) 367–372
FRBR (Functional Requirements for Bibliographical Records) model 63–68, 127–129, 131–136

G
Gender studies, Terminological analysis of 340–345
Geographic location, Influence on research topics 307–312

H
Hermeneutics in knowledge organization 29–34
Hjørland’s theory on knowledge organization 256–260
– Comparison with Feinberg’s theory on subjectivity in classification 256–260
Hjørland’s typology of epistemological positions 242–248
– Applied in analysis of Kaiser’s theory of systematic indexing 243–248
Human clustering of concepts, Comparison with automated clustering 336–338. See also Automated clustering of concepts, Category clusters, Clustering of concepts
Husserl’s phenomenological approach to human experience
– Noesis (component of the approach) 250–255
Hypertext organization of documents 177–183
Hypertext systems, Semantic structure of 177–183

I
Image indexing 77–82. See also Indexing of visual collections, Multilingual image indexing, Multilingual image retrieval
Indexing consistency
– And indexing exhaustivity 144–150
– Measures 144–150
Indexing exhaustivity, And indexing consistency 144–150
Indexing of visual collections 289–293. See also Image indexing
InFile Evaluation Campaign (INformation, FILtering, Evaluation) 370–371. See also Filtering systems, Information filtering in competitive intelligence process
Info-aesthetics 272–274
Information filtering in competitive intelligence process 367–372. See also InFile Evaluation Campaign (INformation, FILTERing, Evaluation)

Information organization frameworks
– Structure of 262–268

Information retrieval
– By children 354–355
– In crosscultural context 91–96
INSPEC Thesaurus 184–189. See also Bilingual thesaurus, Compendex Thesaurus, Thesaurus, Thesaurus construction, Thesaurus development

Institutional identity (Concept), Relationship with national libraries 113–116

Institutional repositories in the UK 196–201

Integrative Level Classification (ILC) 203–209

Interdisciplinary searching behaviour 184–189

Interoperability between conceptual models 124–129

Interpretation of complex concepts by records management experts and novices 295–301

Intertextual semantics 36–41. See also Social semantics

K

Kaiser’s theory of systematic indexing 242–248

Keysigns 282–287. See also American Sign Language (ASL)

Knowledge organization (Concept)
– As a cultural form 269–274
– Philosophical approach 1

Knowledge organization (Discipline), Recent developments in 210–215

Knowledge organization and representation (KOR), Ethics 361–365

Knowledge organization system analysis 242–248

Knowledge representation of cultures 84–89

Knowledge sharing, Impact of technological developments on 313–318

L

LCSH Tree 171–175. See also under Library of Congress Subject Heading (LCSH)

Learning objects
– Classification of 15–21
– Metadata representations of 15–21

Legals resources workflow, Conceptual representation of 63–68

León Manifesto 203–209

Libraries identity 111–116

Libraries websites 105–110

Library of Congress Subject Headings (LCSH). See also LCSH Tree
– Hierarchical structure of 170–176
– Visualization software tool 170–176

Linguistic tools for specialized translation 374–380

Liu Xin 劉欣 (Classificationist) 275–280

Local practices in indexing, In opposition with standard practices 289–293

Localisation of cultural websites 56–61

Logical positivism, Influence on information systems (IS) 29–34

Logocentrism (Derrida’s concept) 226–227

M

Manovich’s theory on culture 269–274
– Concept of info-aesthetics in 272–274

MÁTrIkSz (General Information Retrieval Language Dictionary) 95–96

Medicine 50–54

Memory process 326–330

Mental models of bibliographic universe 131–136

Metadata interoperability 124–129

Metadata schema 124–129


Mixed translations of Dewey Decimal Classification (DDC) 98–103. See also Dewey Decimal Classification (DDC); See also under Facets
– Norwegian model 99
– Swedish model 99–100

Model for Hypertext Organization of Documents (HTXM) 177–183

Mooers’s Method of Descriptors 266–267

Moving Image Collections (MIC) website 56–61

Multi-document summarization 333–339

Multi-viewed knowledge system for administrative document organization 137–142

Multilingual image indexing 77–82. See also Image indexing, Indexing of visual collections, Multilingual image retrieval
Multilingual image retrieval 77–82. See also Image indexing, Indexing of visual collections, Multilingual image indexing
Multilingual knowledge systems, Application of hermeneutics in 32–34
Museums websites 105–110

N
Name authority control in UK repositories 196–201
Names project (UK) 196–201
NanoNetBase, Search queries from 185–189
Natural phenomena classification 235–240
– Planetary classification 239–240
– Saffir-Simpson Hurricane Scale 238–239
Noesis (Husserl’s phenomenological approach component) 250–255
– Applied on mailboxes images analysis 252–253
– Applied on social tagging analysis 253–254
– Definition 250–251
Non-Sense (Concept), In discourse community 302–306. See also Sense (Concept)

O
Objectivity in classification process 22–27
Ontology of legal resources 63–68

P
PBS.org/teachersource 15–21
Personal classification schemes 137–142
Peschl’s Modes of knowing, As learning objects classification structure 18–19
Phenomena classification model 203–209
Philosophy, Relationship with knowledge organization 1
Prelinger Library’s collections classification structure 22–27
Public administrations, E-mail management practices in 347–352
PubMed 320–324

Q
Qualitative digitization, As a document representational practice 111–116

R
Records management 295–301
Reference resources classification 151–155
Research publications
– From international collaboration 310
– In non-US institutions 309–310
– In US-only institutions 308

S
Schellenberg’s models for archives 118–120
Scientific classification 235–240
– Comparison with bibliographical classification 235–236, 240
Search engines, As tools for specialized translation 374–380
Search support systems design 151–155
Searching behaviour 184–189
Semantic structure of hypertext systems 177–183
Semantic Web 222–227
Sense (Concept), In discourse community 302–306. See also Non-Sense (Concept)
Seven Epitomes (Qi lüe 七略) 275–280
– Analysis of classes 276–277
– Origins and history 277–278
Sign language
– Indexing of 284–287
– Metadata 282–287
– On the Web 282–283
Similarity assessment and conceptual coherence 3–7
Simple Knowledge Organization System (SKOS) 36–41
SKOS 36–41
Sloan Digital Sky Survey (SDSS) 307–312
Social bookmarking 320–324
Social sciences theories, Classification of 203–209
Social semantics 229–233. See also Intertextual semantics
Social software/knowledge technologies 216–221
Social tagging 320–324. See also Tagging
– In communities of practice 313–318
– Of Web documents 118–123
Spanish publications, Terminological analysis of 341–343
Speech Act theory 350–351
Standard practices in indexing, In opposition with local indexing practices 289–293
Structure of information organization frameworks 262–268
Subject cataloguing
– Of Web documents 118–123
Subjectivity in classification 22–27, 256–260
T

Tagging 164–169. See also Social tagging
  – Convergence of (Concept) 167–168
  – Stability of (Concept) 165–167
Tags, As search tool 320–324
Taxonomy checklist applied to corporate websites design 157–160
Tecktonik killer community, Use of social tagging in 316–318
Terminological information on the Web 374–380
Text mining systems 307–312
Theories Classification model 203–209
Thesaurus. See also Bilingual thesaurus,
  Compendex Thesaurus, INSPEC Thesaurus, Thesaurus construction, Thesaurus development
  – As tool for search terms mapping 184–189
  – Of water-related topics 190–194
Thesaurus construction, Semi-automatic management of 43–49
Thesaurus development for the Humanities 70–76
Translation
  – Of the Dewey Decimal Classification (DDC) 98–103
  – Use of terminological tools for 374–380
Trust in epistemology (Concept) 217
Trust in social software/knowledge technologies (Concept) 218–219

U

Uncontrolled vocabularies, Use for image indexing 77–82. See also Controlled vocabularies
United Kingdom, Name authority services 196–201
Universal Decimal Classification (UDC) 50–54
Universal Decimal Classification (UDC) Medicine Class, Restructuration of 50–54
Uruguayan publications, Terminological analysis of 341, 345

V

Variations Project (Indiana University) 127–129
Visual collections, Organization of 289–293
VisualLCSH 170–176. See also LCSH Tree
Vocabulary mapping of reference resources 151–155

W

Warrant (Concept) 222–227
  – In Library and Information studies 222–224
  – On the semantic Web 224–227
Warrant in classification (Concept) 235–240
Water-related terminology 190–194
Water resource management in Mexico 190–194
Websites internal organization, Use of facets for 105–110
Wittgenstein’s Language Games theory, As framework for controlled vocabulary development 354–359
Tenth International Conference of The International Society for Knowledge Organization (ISKO). Held at the École de bibliothéconomie et des sciences de l’information, Université de Montréal, Québec, Canada, August 5–8, 2008.

Organizers
ISKO
Université de Montréal
École de bibliothéconomie et des sciences de l’information, Université de Montréal

Co-sponsored by
OCLC
ISKO-NA
MultiTes, Vocabulary Management Software

Conference Chair
Clément Arsenault
École de bibliothéconomie et des sciences de l’information, Université de Montréal

Programme Chair
Joseph T. Tennis
The Information School, University of Washington

Poster Session Chair
Michèle Hudon
École de bibliothéconomie et des sciences de l’information, Université de Montréal

Proceeding Editors
Clément Arsenault
Joseph T. Tennis
Programme Committee

- Jack Andersen, Royal School of Library and Information Science, Denmark
- Tom Baker, Dublin Core Metadata Initiative
- Clare Beghtol, University of Toronto
- Michael Buckland, University of California-Berkeley
- D. Grant Campbell, University of Western Ontario
- Allyson Carlyle, University of Washington
- Melanie Feinberg, University of Washington
- Jonathan Furner, University of California-Los Angeles
- Claudio Gnoli, University of Pavia, Italy
- Rebecca Green, Dewey Decimal Classification, OCLC
- Birger Hjørland, Royal School of Library and Information Science
- Michèle Hudon, Université de Montréal
- Elin K. Jacob, Indiana University
- Traugott Koch, Max Planck Digital Library, Berlin
- Barbara Kwasnik, Syracuse University
- Kathryn La Barre, University of Illinois at Urbana-Champaign
- María López-Huertas, Universidad de Granada
- Jens-Erik Mai, University of Toronto
- José Ángel Martínez Usero, Technosite, ONCE Fondation, Spain
- Ia McIlwaine, University College London
- Eva Méndez Rodríguez, Universidad Carlos III de Madrid
- Shawne Miksa, University of North Texas
- William Moen, University of North Texas
- Widad Mustafa, Université de Lille 3 Marianne
- Marianne Lykke Nielsen, Royal School of Library and Information Science, Denmark
- Peter Ohly, Sozialwissenschaftliche Fachinformation
- John Old, Napier University
- Hope Olson, University of Wisconsin-Milwaukee
- Uta Priss, Napier University
- Jian Qin, Syracuse University
- Ali Shiri, University of Alberta
- Alexander Sigel, Semantic Knowledge Services
- Richard Smiraglia, Long Island University
- Stuart Sutton, University of Washington
- Douglas Tudhope, University of Glamorgan
- Martin van der Walt, Stellenbosch University
- Marcia Zeng, Kent State University