Contents

Article

Yunseon Choi.
A Practical Application of FRBR for Organizing Information in Digital Environments........ 233

Selected Papers from the 8th ISKO-France Conference, 27-28 June 2011, Lille, Université Charles-De-Gaulle Lille 3

Widad Mustafa El Hadi and Clément Arsenault.
Dynamism and Stability in Knowledge Organization: From one Conference to Another: Toronto 2000, Lille 2011................................. 255

Viviane Couzinet.
Knowledge Organization in Information and Communication Sciences, a French Exception? .... 259

Claudio Gnoli.
Metadata About What? Distinguishing Between Ontic, Epistemic, and Documental Dimensions in Knowledge Organization ...................... 268

Philippe Bourdenet.
The Catalog Resisting the Web: An Historical Perspective ............................................... 276

Orélie Desfriches Doria.
The Role of Activities Awareness in Faceted Classification Development............................283

Viviane Clavier and Céline Paganelli.
Including Authorial Stance in the Indexing of Scientific Documents ..................................292

Book Reviews .................................................................300

Norwell, MA: Springer

ISKO News

KO Literature now searchable online ....................................304
A Practical Application of FRBR for Organizing Information in Digital Environments†

Yunseon Choi

Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 501 E. Daniel Street, MC-493, Champaign, IL 61820-6211, USA, <ychoi10@illinois.edu>

Yunseon Choi received her Ph.D. in Library and information Science from the University of Illinois at Urbana-Champaign. She will join the faculty of the Information and Library Science Department at Southern Connecticut State University in August 2012. Her research interests are diverse aspects of information organization including metadata, ontologies, folksonomy, information seeking behaviors, and information architecture. She is particularly interested in social media and information technology in the context of information organization and access in digital environments. Her research focus in information organization is on bringing the user perspective into the organization and retrieval of digital information.


ABSTRACT: This study employs the FRBR (Functional Requirements for Bibliographic Records) conceptual model to provide in-depth investigation on the characteristics of social tags by analyzing the bibliographic attributes of tags that are not limited to subject properties. FRBR describes four different levels of entities (i.e., Work, Expression, Manifestation, and Item), which provide a distinguishing understanding of each entity in the bibliographic universe. In this research, since the scope of data analysis focuses on tags assigned to web documents, consideration on Manifestation and Item has been excluded. Accordingly, only the attributes of Work and Expression entity were investigated in order to map the attributes of tags to attributes defined in those entities. The content analysis on tag attributes was conducted on a total of 113 web documents regarding 11 attribute categories defined by FRBR. The findings identified essential bibliographic attributes of tags and tagging behaviors by subject. The findings showed that concerning specific subject areas, taggers exhibited different tagging behaviors representing distinctive features and tendencies. These results have led to the conclusion that there should be an increased awareness of diverse user needs by subject in terms of the practical implications of metadata generation.

† This paper is derived from the author’s doctoral dissertation “Usefulness of Social Tagging in Organizing and Providing Access to the Web: An Analysis of Indexing Consistency and Quality.” The author is deeply grateful to her dissertation committee—Dr. Linda C. Smith chairperson, Drs. Allen Renear, Miles Efron and John Unsworth.

Received 25 February 2012; Revised 18 April 2012; Accepted 30 April 2012

1.0 Introduction

This study is part of a larger research project that points out major challenging problems with current Knowledge Organization (KO) systems for web resources, such as subject gateways or web directories: 1) the current systems use traditional knowledge organization systems based on controlled vocabulary, which are not very well suited to web resources (Golub 2006; Nowick and Mering 2003; Macgregor and McCulloch 2006), and 2) information is organized by professionals not by users, which means it does not reflect intuitively and instantaneously expressed current user needs (Golub 2006). In order to
explore users’ needs, we examined social tags, which are user-generated uncontrolled vocabulary. As investment in professionally-developed subject gateways and web directories diminishes (support for both BUBL and Intute, examined in this study, is being discontinued), understanding characteristics of social tagging becomes even more critical.

Social tagging has received significant attention since it helps organize contents by collaborative and user-generated tags. Users’ tags reflect their language because social tagging allows users to add their own tags based on their interests. Several researchers have discussed the impact of tagging on retrieval performance on the web (Bao et al. 2007; Choi 2009; Choy and Lui 2006; Golder and Huberman 2006; Heymann, Koutrika and Garcia-Molina 2008; Kipp and Campbell 2010; Sen et al. 2006; Yanbe et al. 2006). Choy and Lui (2006) have applied the statistical tool of Latent Semantic Analysis (LSA) to the evaluation of tag similarity by examining pairs of tags of singular and plural forms, and concluded that collaborative tagging has a great impact on retrieval. Yanbe et al. (2006) have explored an approach to enhancing search by proposing a link-based ranking metric with social tagging data and investigated the utility of social bookmarking systems. Bao et al. (2007) have explored the use of social annotations to improve web search and stated that social annotations could be useful for web search by focusing on two aspects: similarity ranking (between a query and a web page) and static ranking. On the other hand, Choi (2009) has analyzed tags in order to improve web searching by bringing a more accurate user’s perspective into the design of web navigation. In her research, Choi (2009) has provided a new angle for understanding social tags by considering them as “facets.” Kipp and Campbell (2010) also have conducted a study examining whether tags would be useful for information retrieval by limiting the scope of information to scholarly documents such as academic articles at CiteULike and PubMed online journal database. Several studies have explored tags in the context of indexing languages by comparing tags with controlled vocabularies (Good and Tennis 2009; Kipp 2005). On the other hand, Good, Kawas and Wilkinson (2007) have proposed the semantic social tagging application that helps semantic annotations of data in biomedical literatures. Additionally, there have been studies reporting the other aspects of tags such as task and emotion (Kipp 2007a; Neal 2010; Tonkin et al. 2007). There have been also studies on the comparison of users’ tags and professionals or intermediary indexers’ keywords (Kipp 2007b; Choi 2010a, b and c). Kipp (2007b) especially has examined health-related information tags assigned in PubMed articles. She compared tags from users and descriptors from intermediary indexers. Choi (2010a) has focused on bridging the gap of insufficiency of studies on vocabulary analysis by comparing user-generated tags with professionally-generated index terms regarding web resources. The comparison of users’ tags and indexers’ keywords has been promoted by analyzing indexing consistency (Choi 2010b and c). Furthermore, several researchers have discussed the usefulness of social tagging for cataloging and classification by examining the linguistic aspects of user vocabulary (Makani and Spiteri 2010; Spiteri 2007). However, further research is needed to qualitatively as well as quantitatively investigate social tagging and to systematically verify its quality and benefit, which is the first necessary step to utilize social tagging in digital information organization.

To address identified problems with current web organization systems, we aim to investigate whether user-generated tags through social tagging could be used to enhance access to web resources and provide additional access points beyond professionally-generated ones, and whether we could verify the usefulness of social tagging to obtain benefit from it. In this paper, we particularly investigate tag attributes and tagging behaviors. To provide in-depth investigation on the characteristics of tags, we analyze the bibliographic attributes of tags that are not limited to subject properties. Thus, the following research questions are answered: What are features and patterns of social tagging in describing a web document? Do tags have other bibliographic attributes beyond describing subjects or topics of a document?

The process of identifying bibliographic attributes of tags was based on the Functional Requirements for Bibliographic Records (FRBR) model. Because the attributes defined in the FRBR model were derived from “a logical analysis of the data that are typically reflected in bibliographic records” (IFLA 1998), the model supports a more systematic and meticulous analysis of the attributes of tags.

2.0 Background

2.1 Subject gateways as organizing tools for the web

A growing number of web resources have required new tools for organizing and providing more effective access to the web. Subject gateways and web directo-
Subject gateways can be enumerated by the subject categories they cover (University of Kent 2009). For instance, Social Care Online (http://www.scie-socialcareonline.org.uk/) (professional development support portal), SocioSite (http://www.sociosite.net/) (the University of Amsterdam's social science information system), and SWAP (Social Policy and Social Work) (http://www.swap.ac.uk/) (subject portal providing resources to support teachers and lecturers in this subject) are subject gateways that provide resources in social science subjects. For a psychology subject area, there are CogNet (http://cognet.mit.edu) (MIT portal for the brain sciences), PsychNet.UK (http://www.psychnet-uk.com/) (a comprehensive UK gateway to psychology information) and so on. Doctors.net.uk (http://www.doctors.net.uk/) (Peer led internet resource for UK doctors) and HON (Health On the Net) (http://www.hon.ch/) (international Swiss initiative to make quality guidance about medical treatments and health information available to patients and public) are examples for health and medicine subjects. As examples of subject gateways covering various subject areas, there are BUBL Link (http://www.bubl.ac.uk/index.html) and Intute (http://www.intuite.ac.uk/). BUBL describes itself as ‘Free User-Friendly Access to selected internet resources covering all subject areas, with a special focus on Library and Information Science’ (Wikipedia). Intute is a free web service aimed at students, teachers, and researchers in UK further education and higher education (Wikipedia). BUBL offers broad categorization of subjects based on the Dewey Decimal Classification (DDC) scheme (BUBL Link Home). For each subject, subject specialists like librarians work on the maintenance and development of subject categories. However, it has been noted that BUBL is no longer being updated as of April 2011 (BUBL Link Home), as support for BUBL is being discontinued. The selection for inclusion of resources within the Intute collection considers the quality, relevance and provenance of resources (Abbott 2009). It is reported that Intute mainly uses the Universal Decimal Classification (UDC) and DDC for classification and has adapted them for in-house use. Intute subject specialists collaboratively catalog web documents. However, recently it has been noted that Intute is closing after July 2011 (Intute Home), as support for Intute is being discontinued.

2.2 Challenges of controlled vocabulary for the web

For effective indexing and retrieval, the indexing process needs to be controlled by using a so-called controlled vocabulary (Lancaster 1972). Since the 19th century, controlled vocabularies have been developed and used for subject indexing. Lancaster identifies three major manifestations of controlled vocabulary: bibliographic classification schemes, subject heading lists and thesauri.

Controlled vocabulary has many advantages. One of the major advantages of controlled vocabulary is that it can increase the effectiveness of retrieval by providing unambiguous, standard search terms with a control of polysemy, synonymy, and homonymy of the natural language (Golub 2006; Muddamalle 1998). Another benefit from controlled vocabulary is that it improves the matching process with its systematic hierarchies of concepts featuring a variety of relationships like “broader term,” “narrower term,” “related term,” or “see” and “see also” (Golub 2006; Olson and Boll 2001). However, as there are more and more resources available on the web, existing controlled vocabularies have been challenged in their ability to index the range of digital web resources. The challenges of controlled vocabulary for the web can be summarized as follows.

One of the major challenges of controlled vocabulary in the digital environment is the slowness of revision. Indexing web content requires an updatedthesaurus, but usually subjects are rapidly evolving with new terminology, so it is hard to always keep up-to-date vocabulary (Muddamalle 1998). Golub (2006) also addresses “improved currency” and “hospitality for new topics” as new roles which controlled vocabularies need to take. The other problem is that the construction of controlled vocabularies and indexing are labor-intensive and expensive (Fidel 1991; Macgregor and McCulloch 2006). The process of indexing is conducted by professional efforts requiring ex-
pert knowledge (Olson and Boll 2001). Another obstacle of controlled vocabulary is that it has been developed with a focus on physical and traditional library collections. Traditionally, controlled subject headings have been employed for indexing physical resources, so they need to be flexible or expandable in order to encompass web resources (Golub 2006; Nowick and Mering 2003; Macgregor and McCulloch 2006). For instance, *Library of Congress Subject Headings (LCSH)* is designed to describe monographs and serials, so it might not be specific enough for describing web resources (Nowick and Mering 2003). Furthermore, Nicholson et al. (2001) have discussed the problems with controlled vocabularies in indexing for describing online collections by identifying that “they have a lack of, or excessive, specificity in the subject areas.” Last but not least, controlled vocabulary should be comfortable for users to use, and it should be able to meet the users’ interests and their needs (Golub 2006). Golub mentions “intelligibility, intuitiveness, and transparency” as new challenges for controlled vocabulary.

Using free-text or natural language terms is one alternative to resolve identified problems with controlled vocabulary. Advantages of free-text terms are that they require only non-professional knowledge for searching techniques for users and reflect up-to-date vocabulary (Dubois 1987). Social tagging data is one example of natural language terms, that is, uncontrolled vocabulary assigned by users. Social tagging is a promising way to complement the disadvantages of professional indexing because it is low-cost since a great number of users from everywhere contribute to the creation of tags. Thus, users’ tags might be alternate terms with additional entry points of retrieval that are not easily attained using controlled vocabularies (Hayman 2007; Maltby 1975; Quintarelli 2005). Tags are generally much more current than controlled vocabulary because they are constructed in the process of sensemaking, in that users share their experiences in subject terms reflecting their interests in various communities (Smith 2007). Unlike hierarchical structures (broader and narrower terms) of controlled vocabularies, folksonomies are inherently flat, which allows great flexibility in indexing terms. Moreover, as investment in professionally-developed subject gateways and web directories diminishes (support for both BUBL and Intute, examined in this study, is being discontinued), understanding characteristics of social tagging becomes even more critical. In the next section, more details about social tagging and relevant issues will be described.

### 2.3 Social tagging for organizing the web

Social tagging is described as “user-generated keywords” (Trant 2009). Because tags indicate users’ perspectives and descriptions in indexing resources, they have been suggested as a means to improve search and retrieval of resources on the web. The term “social tagging” is frequently associated with the term “folksonomy,” which was coined by Thomas Vander Wal from ‘folk’ and ‘taxonomy’ (Neal 2007). Folksonomy consists of three elements: users, resources to be described, and tags for describing resources (Vander Wal 2005). Vander Wal (2007) describes “folksonomy” as “user-created bottom-up categorical structure development with an emergent thesaurus.” Quintarelli (2005) defines folksonomy as “user-generated classification, emerging through bottom-up consensus.” Examples of folksonomy sites include Flickr, Delicious, and LibraryThing. Social tagging has been popularized by tagging sites such as Flickr, Technorati and Delicious. Delicious is one of the most popular social bookmarking services, allowing users to add or share and organize tags. The site was established as De.li.cio.us by Joshua Schachter in 2003 and acquired by Yahoo! in 2005, and purchased by AVOS Systems on April 27, 2011 (Wikipedia).

Many researchers have suggested that social tagging has potential for user-based indexing (Golder and Huberman 2006; Lin et al. 2006; Tennis 2006). It can be recognized that the participation of users in building controlled vocabulary is being realized in a social tagging environment where users create or generate search keywords based on their intuitive principles. There has been exploratory research investigating tagging as a more accurate description of resources and reflection of more current terminology. Smith (2007) has asserted that tagging is better than subject headings by investigating tags assigned in LibraryThing and the subject headings assigned from LCSH. LibraryThing is a website that allows users to manage a personal catalog with their own books (Wikipedia).

### 3.0 Methodology

#### 3.1 Sampling of web documents

Because this study is part of a larger research project that aims to investigate whether social tagging would enhance access to web resources and provide additional access points beyond those that are professionally-generated, web documents to be analyzed need to be located at a social tagging site as well as profes-
sional indexing sites for comparison. Thus, web documents were randomly sampled when a web document is located at all three web sites, e.g., a social tagging site and two other professional indexing sites. We extracted tags from a social tagging site, Delicious. Delicious has a broad coverage of web resources, not limited to scholarly documents (e.g., journal articles on CiteUlike.org) or specific types of resources (e.g., photos and videos on Flickr). BUBL and Intute were selected as target subject gateways for professional indexing. Both BUBL and Intute cover various subjects and use traditional knowledge organization systems (see table 1). Only if a web document is found at all three locations (BUBL, Intute, and Delicious) were the tags assigned to the document at Delicious extracted.

**Table 1. BUBL vs. Intute.**

<table>
<thead>
<tr>
<th>Site characteristics</th>
<th>BUBL</th>
<th>Intute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>DDC</td>
<td>UDC and DDC</td>
</tr>
<tr>
<td>Keywords</td>
<td>N/A</td>
<td>Controlled: Several thesauri for their subject relevance and comprehensiveness, e.g., SCIE for Social Welfare, the Hasset, IBSS, LIR for Law, and the NLM MeSH headings for Medicine. Uncontrolled: terms from web sites' titles and descriptions Intute indexers provide</td>
</tr>
<tr>
<td>Subjects covered</td>
<td>Various subjects</td>
<td>Various subjects</td>
</tr>
<tr>
<td>Database</td>
<td>Searchable and browsable</td>
<td>Searchable and browsable</td>
</tr>
</tbody>
</table>

Sampling web documents was based on the 10 subject categories (see table 2) BUBL distinctively provides using DDC numbers as top-level categories. Each top-level category is arranged by about 10 second level sub-categories, sometimes more than 10. In order to avoid potential bias in choosing documents at BUBL, a document was first randomly selected from the list of documents associated with a sub-category, and searched in turn at the other two sites, Intute and Delicious. The method of random sampling of documents was based on the True Random Number Generator (www.random.org). If the first document chosen randomly was not found in Intute or Delicious, then the next choice was made randomly until a web document satisfying the selection criteria was found. A total of 113 web documents were randomly selected for samples when choosing one document per sub-category.

**Table 2. BUBL subject categories.**

<table>
<thead>
<tr>
<th>Top Categories</th>
<th>Subjects covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>000 Generalities</td>
<td>Computing, Internet, Libraries, Information Science</td>
</tr>
<tr>
<td>100 Philosophy and psychology</td>
<td>Ethics, Paranormal phenomena</td>
</tr>
<tr>
<td>200 Religion</td>
<td>Bibles, Religions of the world</td>
</tr>
<tr>
<td>300 Social sciences</td>
<td>Sociology, Politics, Economics, Law, Education</td>
</tr>
<tr>
<td>400 Language</td>
<td>Linguistics, Language learning, Specific languages</td>
</tr>
<tr>
<td>500 Science and mathematics</td>
<td>Physics, Chemistry, Earth Sciences, Biology, Zoology</td>
</tr>
<tr>
<td>600 Technology</td>
<td>Medicine, Engineering, Agriculture, Management</td>
</tr>
<tr>
<td>700 The arts</td>
<td>Art, Planning, Architecture, Music, Sport</td>
</tr>
<tr>
<td>800 Literature and rhetoric</td>
<td>Literature of specific languages</td>
</tr>
<tr>
<td>900 Geography and history</td>
<td>Travel, Genealogy, Archaeology</td>
</tr>
</tbody>
</table>

The selection criteria for sampling web documents were as follows:

- Subject categorizations for selecting documents was based on the top-level category at BUBL;
- A web document had to be located at all three web sites, BUBL, Intute, and Delicious; and,
- A web document having more than 50 taggers at Delicious was selected in order to have a sufficient number of taggers for investigating the characteristics of tagging.

### 3.2 Collection of social tags

A Java-based program was written for tag collection and tag pre-processing. Through the Delicious API, the program collected tags in a JSON (JavaScript Object Notation) format (Crockford 2006). For the period from February to March in 2010, Delicious top 20 tags assigned to 113 web documents were collected for analysis. The collected tags were normalized by checking spelling and word forms.
3.3 Data pre-processing

Data pre-processing was conducted for the collected tags to exclude non-English tags or no tags. The collected tags were checked for spelling, acronyms or singular and plural forms. That is, this step included removing misspelled terms and integrating terms which have different forms of words such as noun, adjective, adverb, and gerund.

3.3.1 An exact match between terms

Based on discussion by Lancaster and Smith (1983), we used the following five rules for specifying an exact match between tags:

- Exactly corresponding including singular/plural variations
  - Ex) aurora to auroras, language to languages
- Variant spellings
  - Ex) organization to organisation
- Word forms (adjectival, noun, or verbal forms)
  - Ex) medicine to medical
- Acronyms or abbreviations and full terms
  - Ex) National Center for Biotechnology Information to NCBI, biotechnology to biotech
- Compound terms
  - Ex) human/body to humanbody to human_body to human, body etc.

In terms of tags, Delicious does not have the feature of adding a space between two terms for a compound term, so if there is a dash, slash, or underscore between two terms, or if two terms were found at the same time in the list of tags from a tagger, they were regarded as a compound term. The dragon toolkit (Zhou, Zhang and Hu 2007), which is a WordNet (http://wordnet.princeton.edu/) based lemmatization tool, was used for checking for English words and stemming, which is for merging inflected forms of indexing words. Acronyms were checked in the Acronyms, Initialisms & Abbreviations Dictionary (Reade and Romaniuk 2005).

3.3.2 Term exclusion

Because users at Delicious come from a worldwide audience, they might have different language backgrounds. Thus, if assigned tags are not in English (e.g., in Spanish, Korean, Chinese, etc.), they are excluded from the analysis. Furthermore, we developed a stoplist, which is a list of terms that can be excluded for processing (see Appendix 1). All tags were checked against the stoplist. The stoplist included an explicit list of the terms that Sen et al. (2006) define as subjective and personal tags, because those types of tags are not meaningful for describing subjects of documents. Table 3 provides the three types of tags and their definitions from Sen et al. and the related examples of tags identified.

<table>
<thead>
<tr>
<th>Types of tags</th>
<th>Definitions</th>
<th>Examples of identified tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual tags</td>
<td>“identifies facts about” a resource e.g., people, places, or concepts</td>
<td>government, social-security, finance etc.</td>
</tr>
<tr>
<td>Subjective tags</td>
<td>“express user opinions” related to a resource</td>
<td>good, worth, recommend, toRead, informative etc.</td>
</tr>
<tr>
<td>Personal tags</td>
<td>having “intended audience of tag applied themselves”</td>
<td>myDaughter, forSon, etc.</td>
</tr>
</tbody>
</table>

Table 3.Sen et al. (2006) three types of tags.

3.4 The scope of data analysis using FRBR

The process of identifying bibliographic attributes of tags was based on the FRBR model. Because the attributes defined in the FRBR model were derived from “a logical analysis of the data that are typically reflected in bibliographic records” (IFLA 1998), the model supports a more systematic and meticulous analysis of the attributes of tags.

FRBR is a conceptual model of the “bibliographic universe” (works, texts, editions, documents and the like) that was developed by the International Federation of Library Associations and Institutions (IFLA 1998). It is intended to guide the development of systems for creating and managing bibliographic records. FRBR identifies four “Group 1” entity types (work, expression, manifestation, and item), defines relationships between them (a work is realized through an expression; an expression is embodied in a manifestation; a manifestation is exemplified by an item), and assigns characteristic attributes to each entity. For instance, works have form, expressions may be in a particular language, manifestations may have a typeface, and items may have a provenance. Figure 1 depicts Group 1 entities and relationships between them. The entity work is defined as “A distinct intellectual or artistic creation,” expression as “the intellectual or artistic realization of a work in the form of alphabetic, musical, or choreographic notation, sound, im-
Using the FRBR model, each entity type is assigned a set of attributes. Works have attributes such as title and form; Expressions have a language attribute (translations of the same work are different Expressions); Manifestations have attributes like typeface; and Items have attributes such as condition and location.

In this research, the scope of data analysis focuses on web documents, so consideration of manifestation and item has been excluded. Only the entities Work and Expression were considered and the attributes of both Work and Expression entities were investigated in order to map the attributes of tags to attributes defined for those two entities. Table 4 illustrates the attributes of Work and Expression among FRBR group 1 entities (IFLA 1998). The attributes emphasized in bold face were only included for coding and other attributes were excluded for coding since it was determined that they are not applicable to web documents. Table 5 shows the final list of FRBR attributes for coding and the coding scheme and coding instructions for tag attributes during content analysis are included in Appendix 2. Since each attribute defined by FRBR is assumed to be disjoint (Renear and Choi 2006), this research set up the principle that coding should not overlap.
The content analysis on tag attributes was conducted on a total of 113 web documents regarding 11 attribute categories defined by FRBR (five categories from Work entity and six categories from Expression entity). In order to improve research reliability and objectivity in the analysis of tag attributes, another coder was recruited and the intercoder reliability between two coders was calculated. The recruited coder was a Ph.D. candidate in Library and Information Science. Two coders independently coded tags based on the coding instruction. A sample of coded web document is provided in Appendix 3.

Regarding the sub sample size for the inter-coder reliability test, Wimmer and Dominick (1987) recommend that between 10% and 25% of the data should be investigated to test intercoder reliability. In this research, 25% of the web document collection selected for data analysis is randomly sampled using the True Random Number Generator (www.random.org). For example, under 000 Generalities categories, the number of selected documents was 8, so sub-sample size in this category is 2. Thus, among 113 web documents, 29 web documents are selected for the inter-

<table>
<thead>
<tr>
<th>Entities</th>
<th>Logical attributes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expression</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>presentation technique (cartographic image/object)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>representation of relief (cartographic image/object)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>geodetic, grid, and vertical measurement (cartographic image/object)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>recording technique (remote sensing image)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>special characteristic (remote sensing image)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>technique (graphic or projected image)</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. FRBR Group 1 entities and logical attributes

### 3.3 Intercoder reliability test

The content analysis on tag attributes was conducted on a total of 113 web documents regarding 11 attribute categories defined by FRBR (five categories from Work entity and six categories from Expression entity). In order to improve research reliability and objectivity in the analysis of tag attributes, another coder was recruited and the intercoder reliability between two coders was calculated. The recruited coder was a Ph.D. candidate in Library and Information Science. Two coders independently coded tags based on the coding instruction. A sample of coded web document is provided in Appendix 3.

Regarding the sub sample size for the inter-coder reliability test, Wimmer and Dominick (1987) recommend that between 10% and 25% of the data should be investigated to test intercoder reliability. In this research, 25% of the web document collection selected for data analysis is randomly sampled using the True Random Number Generator (www.random.org). For example, under 000 Generalities categories, the number of selected documents was 8, so sub-sample size in this category is 2. Thus, among 113 web documents, 29 web documents are selected for the inter-

<table>
<thead>
<tr>
<th>Entities</th>
<th>Logical attributes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work</td>
<td>title of the work (WT)</td>
<td>The title of the work is the word, phrase, or group of characters naming the work. There may be one or more titles associated with a work.</td>
</tr>
<tr>
<td></td>
<td>form of work (WF)</td>
<td>The form of work is the class to which the work belongs (e.g., novel, play, poem, essay, biography, symphony, concerto, sonata, map, drawing, painting, photograph, etc.).</td>
</tr>
<tr>
<td></td>
<td>date of the work (WD)</td>
<td>The date of the work is the date (normally the year) the work was originally created. The date may be a single date or a range of dates. In the absence of an ascertainable date of creation, the date of the work may be associated with the date of its first publication or release.</td>
</tr>
<tr>
<td></td>
<td>intended audience (WI)</td>
<td>The intended audience of the work is the class of user for which the work is intended, as defined by age group (e.g., children, young adults, adults, etc.), educational level (e.g., primary, secondary, etc.), or other categorization.</td>
</tr>
<tr>
<td></td>
<td>context for the work (WC)</td>
<td>Context is the historical, social, intellectual, artistic, or other context within which the work was originally conceived (e.g., the 17th century restoration of the monarchy in England, the aesthetic movement of the late 19th century, etc.).</td>
</tr>
<tr>
<td>Expression</td>
<td>form (EF)</td>
<td>The form of expression is the means by which the work is realized (e.g., through alphanumeric notation, musical notation, spoken word, musical sound, cartographic image, photographic image, sculpture, dance, mime, etc.).</td>
</tr>
<tr>
<td></td>
<td>date (ED)</td>
<td>The date of expression is the date the expression was created (e.g., the date the particular text of a work was written or revised, the date a song was performed, etc.). The date may be a single date or a range of dates. In the absence of an ascertainable date of expression, the date of the expression may be associated with the date of its publication or release.</td>
</tr>
<tr>
<td></td>
<td>language of expression (EL)</td>
<td>The language of the expression is the language in which the work is expressed. The language of the expression may comprise a number of languages, each pertaining to an individual component of the expression.</td>
</tr>
<tr>
<td></td>
<td>summarization of content (ES)</td>
<td>A summarization of the content of an expression is an abstract, summary, synopsis, etc., or a list of chapter headings, songs, parts, etc. included in the expression.</td>
</tr>
<tr>
<td></td>
<td>use restrictions on the expression (EU)</td>
<td>Use restrictions are restrictions on access to and use of an expression. Use restrictions may be based in copyright, or they may extend beyond the protections guaranteed in law to the owner of the copyright.</td>
</tr>
<tr>
<td></td>
<td>technique (graphic or projected image) (ET)</td>
<td>Technique is the method used to create a graphic image (e.g., engraving, etc.) or to realize motion in a projected image (e.g., animation, live action, computer generation, 3D, etc.).</td>
</tr>
</tbody>
</table>

Table 5. FRBR attributes and description (IFLA Study Group 1998).
coder reliability test (Table 6). Thus, among 1,879 tags assigned to 113 documents, 442 tags assigned to 29 web documents are coded for intercoder reliability.

<table>
<thead>
<tr>
<th>Top Categories</th>
<th>The number of selected documents</th>
<th>25%</th>
<th>The number of documents for intercoder reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>000 Generalities</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>100 Philosophy and psychology</td>
<td>6</td>
<td>1.5</td>
<td>2</td>
</tr>
<tr>
<td>200 Religion</td>
<td>12</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>300 Social sciences</td>
<td>12</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>400 Language</td>
<td>9</td>
<td>2.25</td>
<td>2</td>
</tr>
<tr>
<td>500 Science and mathematics</td>
<td>10</td>
<td>2.5</td>
<td>3</td>
</tr>
<tr>
<td>600 Technology</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>700 The arts</td>
<td>21</td>
<td>5.25</td>
<td>5</td>
</tr>
<tr>
<td>800 Literature and rhetoric</td>
<td>15</td>
<td>3.75</td>
<td>4</td>
</tr>
<tr>
<td>900 Geography and history</td>
<td>12</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>113</td>
<td>28.25</td>
<td>29</td>
</tr>
</tbody>
</table>

Table 6. The number of documents for intercoder reliability test.

There are a number of measures of intercoder reliability. Lombard, Synder-Duch and Bracken (2005) describe several measures commonly used in social science and communication such as percent agreement, Holsti’s method, Scott’s pi (Π), Cohen’s kappa (κ), and Krippendorff’s alpha (α). The percent agreement index has advantages of simplicity and ease of calculation, but it records only agreements and disagreements. This index also has a flaw in that it does not account for agreement occurring by chance. Holsti’s method (1969) is a variation on the percent agreement index; it accounts for the situation in which the coders evaluate different units. But, when two coders evaluate the same units, the results by Holsti’s method are the same as those by the percentage agreement index of reliability because it calculates percent agreement between two coders (Hayes 2007; Lombard, Snyder-Duch and Bracken 2005). Scott’s pi (1955) takes into account both the observed proportion of agreement and the proportion that would be expected by chance. Yet, Scott’s pi has a limitation to two coders and nominal data (Hayes 2007). On the other hand, several researchers (Bakeman 2000; Dewey 1983) recommend Cohen’s kappa (κ), one of the widely used measures for intercoder reliability. Cohen’s kappa is identical to Scott’s pi in that it accounts for agreement expected by chance. The equation for kappa (κ) is as follows:

\[
\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}
\]

Pr(a): agreement, observed
Pr(e): agreement, expected by chance

Unlike Scott’s pi, the assumption of kappa is that the same two coders have coded all units, so it cannot be applicable to situations where different pairs of coders have coded different subsets of the units (Craig 1981). Krippendorff (1978, 1987, 2004) also criticizes that Cohen’s kappa (κ) is not appropriate for testing intercoder agreement. Krippendorff insists that because Cohen’s kappa (κ) defines chance as “the statistical independence of two coders’ use of categories,” the categories one coder uses are not predictable from the categories the other coder uses.

Krippendorff’s alpha (α) (1980) is also a commonly used measure for intercoder reliability. It is considered to be very flexible as it can account for different sample sizes and missing data, and can be applied to any number of observers, any number of categories, and any level of measurements, e.g., nominal, ordinal, interval, ratio, and more (Hayes 2007; Lombard, Snyder-Duch and Bracken 2005; Krippendorff 2004). Alpha (α)’s general form is as follows (Krippendorff 2004):

\[
\alpha = 1 - \frac{D_o}{D_e}
\]

D_o: disagreement, observed
D_e: disagreement, expected by chance

α = 1 means observers agree perfectly, i.e., perfect reliability and the value of D_o is zero. Also, α = 0 means the absence of reliability, and D_e=D_o. Thus, α’s range is explained by:

\[
1 \geq \alpha \geq 0 \quad \{ \begin{array}{l} \text{Systematic disagreement} \\
\pm \text{Sampling errors} \end{array} \}
\]

Although many reliability measures have been used and discussed by several researchers, there has been no consensus on a best measure for reliability, and each index has its own qualities and assumptions (Lombard
et al., 2005; Taylor & Watkinson, 2007). In this research, therefore, four indices mentioned above, i.e., Holsti’s method, Scott’s pi (II), Cohen’s kappa (κ) and Krippendorff’s alpha (α), are used to test intercoder reliability. Calculating and reporting reliability by using more than one index is a preferred approach that can take into account any bias or weaknesses caused by the results from one (Lombard, Snyder-Duch and Bracken 2005).

4.0 Results

The results of the analyses of tag attributes based on the FRBR model illustrated important tag attributes and tagging behaviors by subject.

4.1 Results of the intercoder reliability test

The intercoder reliability test was calculated by using the Holsti method, Scott’s pi, Cohen’s kappa and Krippendorff’s alpha. In terms of criteria for acceptability, index scales are analogous but it has been cautioned that different indices measure different things (Lombard, Snyder-Duch and Bracken 2005; Neuen-dorf 2002). Therefore, a satisfactory level depends on the index used (Taylor and Watkinson 2007). Holsti (1969) suggests the agreement level of 85 % or more for the acceptable level. Banerjee et al. (1999) suggest that Cohen’s kappa levels should exceed 0.75 for excellent agreement beyond chance, between 0.40-0.70 is fair to good agreement beyond chance, and <0.40 is poor agreement. Landis and Koch (1977) have provided a more detailed list of interpretation of kappa: 0.81 – 1.00 is almost perfect agreement, 0.61 – 0.80 is substantial agreement, 0.41 – 0.60 is moderate agreement, 0.21 – 0.40 is fair agreement, 0.0 – 0.20 is slight agreement and < 0 is poor agreement. For the case of Krippendorff’s alpha, it has been suggested to exceed 0.70 for excellent agreement (Krippendorff 2004; Taylor and Watkinson 2007). In this research, in four indices, the results of the intercoder reliability test showed an excellent agreement as shown in Table 7.

<table>
<thead>
<tr>
<th>Measure of reliability</th>
<th>Value</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holsti</td>
<td>.8824</td>
<td>442</td>
</tr>
<tr>
<td>Scott’s pi</td>
<td>.7963</td>
<td></td>
</tr>
<tr>
<td>Cohen’s kappa</td>
<td>.7963</td>
<td></td>
</tr>
<tr>
<td>Krippendorff’s Alpha</td>
<td>.7965</td>
<td></td>
</tr>
</tbody>
</table>

Table 7. Results of intercoder reliability test using four indices.

In order to investigate the degree of reliability among subject areas, the reliability test on each subject area was performed. The results of intercoder reliability test using four indices demonstrated that the Literature subject showed the lowest level of agreement among 10 different subject areas (Figure 2).

Table 8 illustrates the cross-tabulation of coded data by two coders on the Literature subject. It was found that there was especially low agreement between two coders on two attribute categories, i.e., WF (Form of Work entity) and EF (Form of Expression entity). The examples of those tags were Books, Database, Magazine, Journal, and Encyclopedia. This disagreement on those attributes was caused by the fact that the documents, tagged with a term “Book,” include the list of books or provide a feature of searching for books rather than books themselves (see Table 9). However, current definitions provided

![Four indices for intercoder reliability](image-url)
by FRBR do not explicitly distinguish these two attributes (i.e., WF and EF) about web documents. To make FRBR more applicable in practical terms, FRBR should be able to describe digital heterogeneous media resources that are available in various formats and attributes.

As discussed above, the results of the intercoder reliability test (see Table 7) were very satisfactory with excellent agreement for all four indices (Banerjee et al. 1999; Holsti 1969; Krippendorff 2004; Landis and Koch 1977; Taylor and Watkinson 2007), but it is very important to note that reliability and validity are different. Reliability is concerned with the consistency of the measurement while validity is related to the strengths of the results. Krippendorff (2008, 357) asserts that validity is about truth and reliability relates to trust. He also argues that “reliability cannot guarantee validity.” Thus, the results of the intercoder reliability test do not determine the validity of the conclusions on tag analysis, but instead, they contribute to enhancing confidence in reliability. In the following sections, the results on the analysis of tag attributes are discussed for the whole collection of documents.

4.2 Categories of tag attributes

During the process of content analysis on tag attributes, if a tag was determined to be a term related to subjects or topics describing documents, the tag was categorized as “Subject.” Also, if a tag was identified as a term that cannot be categorized into any of the categories defined by FRBR, the tag was categorized as “Others.” Finally it was determined that the tags included in the “Others” would be assigned to subcategories such as Feature, Utilization, and Institution etc, and the discussion of those tags will be provided later. The findings on the analysis of tag attributes are depicted in Figure 3. Figure 3 illustrates that among tags assigned to the sampled documents, in the pie chart, 26% of tags were subject-related terms, 27% of tags were matched into the attributes of FRBR, and 47% of tags were categorized into other attributes. This illustrates that many tags (about 74%) include additional properties beyond subject or topic terms.
4.3 Tagging behaviors

In order to investigate whether the attributes of tags could be described by the FRBR attributes, a matching process was conducted between tags and FRBR attributes. Tags were identified based on the attribute categories defined by FRBR as shown in Table 10. Table 10 excludes the WT (Title of work entity) category where tags consist of terms used in the title of the document. Regarding the tags related to subject terms, in Language, Literature, and Geography subject, the number of subject-related tags was relatively low (Figure 4 and Figure 5).

Figures 6-8 below illustrate that in terms of web documents in those three subjects, taggers tended to focus more on other properties of documents rather than the subjects or topics of documents, that is, the Form of Work entity (WF) and Form of Expression entity (EF). Since the figures mainly show the comparison of subject-related tags and FRBR categorized tags, the “Others” category is not represented in those figures. A more in-depth analysis was conducted on the tendency of tagging in terms of 11 FRBR attribute categories. Figures 9 and 10 demonstrate that taggers tend to mainly assign tags on attributes related to WT (Title attribute of FRBR Work entity) and WF.

In order to investigate the features and patterns of social tagging in assigning attributes matching those defined in FRBR, a thorough examination was conducted on tags categorized by FRBR attributes. Figure 11 and 12 show tag frequency on the categories defined by FRBR in terms of 10 different subject areas.

![Figure 3. Tag frequency and attribute categories](image)

<table>
<thead>
<tr>
<th>Entities</th>
<th>Attributes</th>
<th>Identified tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work</td>
<td>Form of work (WF)</td>
<td>reference, journal, research, magazine, news, paper, article, dictionary, archive, database, directory, book, essay, scripture, gov-doc, encyclopedia, glossary, tutorial</td>
</tr>
<tr>
<td></td>
<td>Date of work (WD)</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Intended audience (WT)</td>
<td>baby, doctor, engineer, artist, dealer, architect, author, writer, children, illustrator, poet, teacher</td>
</tr>
<tr>
<td></td>
<td>Context for the work (WC)</td>
<td>world, war, uk, primary source, 18c, India, usa, middleeast, federal, Boccaccio, Medieval, ancient</td>
</tr>
<tr>
<td>Expression</td>
<td>Form (EF)</td>
<td>music, ebook, texts, iconography, images, statistics, word, video, vocabulary, etext, bibtex, pictures, photos, multimedia, graphic, audio, sound, illustration, posters</td>
</tr>
<tr>
<td></td>
<td>Date (ED)</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Language of expression (EL)</td>
<td>English, Hebrew, Greek,</td>
</tr>
<tr>
<td></td>
<td>Summarization of content (ES)</td>
<td>list</td>
</tr>
<tr>
<td></td>
<td>Use restrictions on the expression (EU)</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Technique (graphic or projected image) (ET)</td>
<td>graphic organizer, flash</td>
</tr>
</tbody>
</table>

*Table 10. Identified tags and related FRBR attributes*
4.4 FRBR Intended audience of Work entity (WI)

As shown in Figure 11 and 12, the tag frequency on FRBR attributes formed a different tendency depending on subject categories. For example, in three subject areas, Technology, Arts and Literature subjects, the tag frequency on FRBR WI (intended audience) attribute was relatively high (see Figure 13), which means that taggers tend to consider audience in these subject areas. In the Technology subject, the tags applied to the WI category were doctor, engineer etc. On the other hand, in the Art subject, the tags were artists, architects, and dealers etc. In the Literature subject, the tags were author, poet, children, and writer etc. It can be inferred that high frequency on the WI category in those subject areas reflects the characteristics of different user needs for metadata. For example, in Literature, many documents are intended for adults, so if a document is related to resources for children, taggers tend to specifically indicate it by assigning a tag, “children” as the intended audience.

4.5 FRBR Form of Expression entity (EF)

In terms of Natural sciences and Geography, the findings on tag frequency of the EF category showed relatively high proportions (respectively, 21% and 28%) in comparison with those of other subject categories (Figure 14). In both subject areas, the tags assigned to the EF category were image, video, picture, and photos etc. It implies that web documents in Natural sciences and Geography are mainly characterized by taggers with focus on specific forms.

4.6 Other tag attributes

Besides the categories mentioned above, the proportion of tags having other types of attributes was 47% (Figure 15). Concerning the other attributes of tags that were not categorized into any attribute categories (FRBR attributes and subject categories), three subcategories were developed to sort out those tags, i.e., Feature, Utilization, and Institution. Also, if a tag could not be assigned to any of the subcategories
Figure 6. Tag frequency rates on Language subject.

Figure 7. Tag frequency rates on Literature subject.

Figure 8. Tag frequency rates on Geography subject.
Figure 9. Tag frequency on FRBR attributes (bar graph)

Tag Frequency on FRBR Attributes

Figure 10. Tag frequency on FRBR attributes (pie chart)

Figure 11. Tag Frequency on FRBR attributes over all subjects (bar graph)
Figure 12. Tag Frequency on FRBR attributes over all subjects (bar graph II)

Figure 13. Tags on intended audience (WI)

Figure 14. Tags on forms of Expression (EF)
mentioned above, the tag was labeled as “Not Applicable” (Table 11). The tags in the Utilization subcategory show rather subjective or personal properties. Those tags such as resources, learning, teaching, and job imply a user’s intent to use documents for particular purposes.

5.0 Limitations

We limited the scope of sample web documents to the common document collection of BUBL and Intute, and only if a web document was listed at both locations were tags assigned to the web document at Delicious collected and analyzed. Thus, conclusions about properties of tags in Delicious were limited to web documents selected for inclusion in subject gateways and indexed by professional indexers. In addition, analysis for content analysis of tag attributes focused on the top 20 ranked tags. A more thorough study of tagging behavior would encompass a larger number of assigned tags associated with each document.

6.0 Conclusion

In order to characterize the features and patterns of tags, the content analysis of tag attributes was performed based on attributes defined by the FRBR model. The findings identified the bibliographic attributes of tags beyond describing subjects or topics of a document. The findings also showed that tags have essential attributes matching those defined in FRBR. In terms of FRBR attributes, the results showed that taggers tend to mainly assign tags on attributes related to WT (Title attribute of FRBR Work entity) and WF (Form attribute of FRBR Work entity). Furthermore, in terms of specific subject areas, taggers exhibited different tagging behaviors representing distinctive features and tendencies. For three subject areas, Technology, Arts and Literature subjects, tag frequency on the FRBR WI (intended audience) attribute was relatively high, which means that taggers tend to consider audience in these subject areas. In terms of Natural sciences and Geography, the

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature</td>
<td>Feature is a technical feature about web documents. It reflects the characteristics of web documents.</td>
<td>academic, library, conference, community, search, online, bookmarkbar, open_access, web2.0, library2.0, homepage, networking, links, blog, tools, access, browse, portal, community, forum, public-domain, wiki</td>
</tr>
<tr>
<td>Utilization</td>
<td>Utilization is about the implied purpose of usage.</td>
<td>resources, education, information, learning, e-learning, writing, reading, study, teaching, job, career, tutorial</td>
</tr>
<tr>
<td>Institution</td>
<td>Institution</td>
<td>Association, organization, foundation</td>
</tr>
<tr>
<td>Not Applicable</td>
<td>cannot be determined as any categories above</td>
<td>imported, flickr</td>
</tr>
</tbody>
</table>

Table 11. Tag categories for other attributes
tag frequency of EF (Form attribute of Expression entity) category showed relatively high proportion in comparison with those of other subject categories. This indicated that web documents in both those subject areas were characterized by taggers with a focus on specific forms. The other attributes of tags were sorted into three sub categories, Feature, Utilization, and Institution. These results have led to the conclusion that there should be an increased awareness of diverse user needs by subject in order to improve metadata in practical applications.

It should be noted that since the scope of data analysis focuses on tags describing web documents, in this research, consideration of the FRBR Manifestation entity and Item entity has been excluded. Given the characteristics of web documents in terms of “web publishing,” a web document can be viewed as the “digital embodiment” of a print book or a print journal. In that case, FRBR definitions of manifestation also needed to be extended to identify different manifestations with the same content.

The results found in this research revealed that while conducting content analysis of tag attributes, there was some disagreement between two coders on two FRBR attribute categories, i.e., WF (Form of Work entity) and EF (Form of Expression entity). The examples of those tags were Books, Database, Magazine, Journal, and Encyclopedia. This disagreement on those attributes was caused by the fact that the documents, tagged with a term “Book,” include the list of books or provide a feature of searching for books rather than books themselves. However, current definitions provided by FRBR do not explicitly distinguish these two attributes about web documents. To make FRBR more applicable, FRBR should be able to describe digital heterogeneous media resources which are available in various formats and multi-dimensional structures. Therefore, an important future direction for my research will involve expanding current FRBR definitions on entities and attributes for web documents in digital environments.

References

Choi, Yunseon. 2010a. "Enhancing access to the Web: vocabulary analysis on users’ tags and professionals’ index terms." Poster presented at iConference 2010, February 3-6, Champaign, Illinois.
Craig, Robert T. 1981. Generalization of Scott’s Index of Intercoder Agreement. Public opinion quarterly 45: 260-64.


Hayes, Andrew F. 2007. Answering the call for a standard reliability measure for coding data. Communication methods and measures 1: 77-89.


Tonkin, Emma, Baptista, Ana Alice, Hooland, Seth van, Resmini, Andrea, Eva, Mendéz, and Neville, Liddy. 2007. Kinds of tags: a collaborative research study on tag usage and structure. *The 6th European


### Appendix 1. Stoplist

<table>
<thead>
<tr>
<th>affordable</th>
<th>informative</th>
</tr>
</thead>
<tbody>
<tr>
<td>awesome</td>
<td>personal</td>
</tr>
<tr>
<td>babyas</td>
<td>popular</td>
</tr>
<tr>
<td>bad</td>
<td>portal</td>
</tr>
<tr>
<td>base</td>
<td>post_graduate</td>
</tr>
<tr>
<td>befolkning</td>
<td>postgraduate</td>
</tr>
<tr>
<td>best_of_the_web</td>
<td>pre-kindergarten</td>
</tr>
<tr>
<td>bestoftheweb</td>
<td>professional</td>
</tr>
<tr>
<td>bookmarksbar</td>
<td>professional_resource</td>
</tr>
<tr>
<td>bourse</td>
<td>read_later</td>
</tr>
<tr>
<td>by</td>
<td>recommend</td>
</tr>
<tr>
<td>ccstuff</td>
<td>recommended_site</td>
</tr>
<tr>
<td>cdweb</td>
<td>recommendedsite</td>
</tr>
<tr>
<td>check</td>
<td>ref_source</td>
</tr>
<tr>
<td>collectibles</td>
<td>search</td>
</tr>
<tr>
<td>convenient</td>
<td>self-help</td>
</tr>
<tr>
<td>cool</td>
<td>sharing</td>
</tr>
<tr>
<td>download</td>
<td>staring_site</td>
</tr>
<tr>
<td>fact</td>
<td>startingSite</td>
</tr>
<tr>
<td>favorite</td>
<td>student</td>
</tr>
<tr>
<td>for_student</td>
<td>stumbleupon</td>
</tr>
<tr>
<td>free</td>
<td>stumbleuponfavorite</td>
</tr>
<tr>
<td>free.to.everyone</td>
<td>tip</td>
</tr>
<tr>
<td>funny</td>
<td>to.read</td>
</tr>
<tr>
<td>good</td>
<td>toread</td>
</tr>
<tr>
<td>good_info</td>
<td>to_be_better_tagged</td>
</tr>
<tr>
<td>good_information</td>
<td>toblog</td>
</tr>
<tr>
<td>good_practice</td>
<td>tocatalog</td>
</tr>
<tr>
<td>gooddesign</td>
<td>todescribe</td>
</tr>
<tr>
<td>good-design</td>
<td>toread</td>
</tr>
<tr>
<td>goodinfo</td>
<td>useful</td>
</tr>
<tr>
<td>goodpractice</td>
<td>useful_link</td>
</tr>
<tr>
<td>grad</td>
<td>useful_stuff</td>
</tr>
<tr>
<td>grad_school</td>
<td>usefulstuff</td>
</tr>
<tr>
<td>gradschool</td>
<td>vital_record</td>
</tr>
<tr>
<td>guide</td>
<td>vitalrecord</td>
</tr>
<tr>
<td>help</td>
<td>worth</td>
</tr>
<tr>
<td>how_things_work</td>
<td>wow</td>
</tr>
<tr>
<td>howto</td>
<td></td>
</tr>
<tr>
<td>humor</td>
<td></td>
</tr>
<tr>
<td>interesting</td>
<td></td>
</tr>
</tbody>
</table>
Appendix 2. Coding Instruction

If you determine that a tag can be associated with a specific category of FRBR attributes, enter a number “1” in the cell. If you determine that a tag cannot be associated with any categories of FRBR attributes, leave the cell blank, and you can put your comments in the “Notes” cell, if possible. For instance, if you determine that a tag can be regarded as a “subject term”, enter an “S” in the Notes cell. Otherwise, describe it, if possible, or just put a question mark “?”. 

<table>
<thead>
<tr>
<th>Subject</th>
<th>Title</th>
<th>Tags</th>
<th>Work</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>Institute for Psychohistory: <a href="http://www.psychohistory.com/">http://www.psychohistory.com/</a></td>
<td>psychology, history, politics, psychohistory, science, culture, reference, world, war, abuse, theory, academic, sociology, parenting</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Appendix 3. A sample of coded web document based on FRBR attributes

<table>
<thead>
<tr>
<th>Subject</th>
<th>Title</th>
<th>Tags</th>
<th>Work</th>
<th>Expression</th>
<th>Not Applicable</th>
</tr>
</thead>
<tbody>
<tr>
<td>890 Poetry, general resources</td>
<td>Modern Haiku, <a href="http://www.modernhaiku.org/">http://www.modernhaiku.org/</a></td>
<td>haiku, poetry/poems, japan, literature, magazine, writing, journal, words, review, world, creative writing, online</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dynamism and Stability in Knowledge Organization: From one Conference to Another: Toronto 2000, Lille 2011

Widad Mustafa El Hadi* and Clément Arsenault**

*Charles de Gaulle University Lille 3; UFR IDIST, Domaine du Pont de Bois, Villeneuve d’Ascq 59653, France < widad.mustafa@free.fr>

**EBSI, Université de Montréal, C-2012 Lionel-Groulx, C.P . 6128, succursale Centre-ville, Montréal (Québec) CANADA H3C 3J7 <clement.arsenault@umontreal.ca>

Since its creation in 1996, the French chapter of ISKO has been concerned with knowledge organization issues. This topic has been dealt with from different angles: knowledge organization structures, tools for mediation, forms and mechanisms for knowledge sharing. Given that these issues are at the center of information production and access as well as knowledge dissemination, the 8th edition of the ISKO-France conference aimed to focus specifically on the subject of stability and dynamism in the concepts and paradigms underlying knowledge organization research. Eleven years after the sixth International ISKO Conference organized by the Faculty of Information Studies at the University of Toronto on this very theme, the French ISKO chapter,¹ convinced of the importance of this theme, proposed to revisit it.

Access to information and thereby to knowledge is an important social, political, cultural and economic stake. This demands that we take a reflexive look at the theories, paradigms and concepts underlying the organization and the circulation of information and knowledge. The past years have witnessed an increase in the potentials of information technology. New socio-technical practices have emerged. This 2011 edition aimed in particular to examine the mutations that modes and structures of knowledge organization might have undergone in the face of technological advances driven by the Web, especially by the Social Web. Conversely, we also sought to ascertain which modes and structures had resisted the technological upheaval provoked by the Web. What are the reasons for their stability or dynamism? What is the impact of the societal mutations induced by the penetration of the Web in every aspect of scientific and professional activity on modes of knowledge organization and on modalities for the production and
circulation of knowledge? What is or will be the impact of the so-called “Semantic Web” on knowledge organization research? What repercussions may we expect on models, structures and knowledge representation modes? Are we in the face of an evolution or a revolution, a break-off or continuity?

The call for papers identified four main topics:

1. Historical and epistemological foundations of knowledge organization (KO)
   - paradigms underlying research in KO
   - structures and relations in KO
   - role and influence of culture for the design of KO tools
   - evolution and stability of KO paradigms in different societies or cultures

2. Knowledge organization systems
   - evolution of controlled vocabularies (classification languages, thesauri,..) in the semantic web era
   - interoperability of systems and devices for KO
evolution of systems for KO and the social collaborative practices of web 2.0

3. Mutations in professional practices
   - transformation of documentary forms
   - collaborative practices in documentation services
   - interoperability of knowledge representation tools
   - metamorphosis of the documentation space in its relation with sense making, with form and time
   - new editorial forms, new representation modes for information and knowledge circulation

4. Uses and users of knowledge organization systems
   - users practices of KO systems
   - uses of new tools and devices for classification and categorization (folksonomies, folksontologies)

Our comparison between the two conferences was based on two elements—first, the contributions to ISKO-Lille (2 keynote addresses, 26 papers and 7 posters) and the contributions to the Sixth ISKO International Conference held in Toronto, and second, the panel addressing the general theme of the conference organized by Clément Arsenault and Widad Mustafa El Hadi. The invited panelists were: Charles Huot, TEMIS, France; Michèle Hudon, EBSI, Université de Montréal; Sylvie Leleu-Merviel, U. de Valenciennes; Maria López-Huertas, Facultad de Biblioteconomía y Documentación, Université de Grenade (Spain); Peter Ohly, President, ISKO International; Fabrice Papy, Université de Nancy 2; Barbara Sosinska-Kalata, Université de Varsovie; Manuel Zacklad, DICEN-CNAM, Paris.

We asked the panelists to report either on theoretical work linked to stability or dynamism of theories, concepts and paradigms in the field of knowledge organization or to focus on innovative applications but with an emphasis on the theoretical and epistemological underpinnings of these practical works. We suggested that their contributions and remarks could focus on elements pertaining to their expertise in knowledge organization or grounded on the elements drawn from the abstracts of accepted proposals for ISKO 2011 in Lille.

Frequency analysis performed on uniterms and phrases from titles and abstracts of papers presented at the two ISKO conferences (Toronto 2000 and Lille 2011), helped to highlight the differences and similarities between these two meetings. The results of this analysis were presented as word clouds to panelists, whom we asked to participate in the closing session of the conference in Lille, in the hope that these patterns of representation might provide food for thought. These were simply intended as a guide and could be used a starting point for their analysis. The main questions raised were partially drawn from our Call for Papers. Although rudimentary, this statistical exercise revealed quite clearly, however, what were the main themes discussed during the two conferences and showed the progress over the last decade. The evolution of themes reflects the dynamism of our community and of our field of study. The frequency analysis also revealed what are the recurring themes common to both conferences, unwavering and proud representatives of the stability which also characterizes our field.

Unsurprisingly, key themes such as indexing, classification, semantic analysis, information and knowledge are present in both 2000 and 2011. The analysis shows however that the attendants were more concerned ten years ago with the foundations and the theoretical basis than is the case today, at least for this conference. In 2011, in terms of classification, it is not so much the foundations and the systems that were the subject of the studies presented but rather the practical uses and the integration of classification systems within larger information systems. Concerning indexing, we notice that we have discussed the linguistic foundations less than a decade ago. Instead, at the 2011 conference, many texts addressing the social use of indexing and its integration into the information practices of users were presented. New practices such as collaborative tagging and the products of social indexing, such as folksonomies, hardly known ten years ago, are now brought to the forefront as a topic of
study by several researchers. In 2011 the web is ubiquitous. There is no mention even of the Internet, which has become an implicit element of the world in which we evolve. The glut of information available on the web has resurrected, from a new perspective, the issue of archiving, including systems of open archives, a theme that was absent from the conference of 2000.

The most apparent difference between the topics discussed during the two conferences is the clear emergence of everything related to the social web. The terms “collective,” “collaborative,” “community,” and “interoperability,” almost absent in 2000, are, in 2011, among those with the highest frequency of occurrence. This is not surprising considering the enormous progress made during the last decade in the democratization of information achieved through the spectacular development of the web and its integration in all spheres of society. The appropriation of information systems by non-expert user communities requires that we devote more and more studies toward users and refocus these on their information needs. This theme appears very clearly in 2011 while in 2000 it was all but emerging. A decade ago, the emphasis was clearly mostly on systems, design, interfaces, terminology and thesauri, whereas today, studies are turning their attention to social uses and behavior of new classes of users.

The two-day conference in Lille had an outstanding selection of speakers from the domains of knowledge organization, bibliographic classification, Information and Communication Sciences. The conference attracted researchers representing several of the major Information and Communication Departments in France, along with Library and Information Science Schools and industrial firms from many countries: Belgium, Canada, France, Germany, Italy, Poland, Saudi Arabia and Spain. The conference took place at the University of Lille 3 and the program included two keynote addresses, 26 papers organized into 9 sessions, a poster session and a panel addressing the general theme of the conference. The conference proceedings were published by Hermès-Science, Traité des Sciences et Techniques de l’Information, Collection Organisation de l’information (El Hadi andArsenault 2012).

The articles presented in this issue of Knowledge Organization (vol. 39, no. 4), are translated and updated versions of the presentations published in the conference proceedings. The first two articles are the two keynote addresses: “Knowledge organization in the context of Information and Communication Science: A French exception?” by Viviane Couzinnet, Université de Toulouse 3. The second keynote address, “Metadata about what? Distinguishing between ontic, epistemic, and documental dimensions in knowledge organization,” by Claudio Gnoli, Università di Pavia. Three other articles, representative of the French research in KO and Information and Communication Sciences, have also been selected. We hope that presenting the French Chapter and its work to the international scientific community will allow a better understanding and appreciation of the French researchers’ endeavor and of their contributions to the field of Knowledge Organization.

Note

1. The French Chapter was founded in December 1996 at the initiative of Jacques Maniez and Danièle Degez, both ISKO members at that time. Jacques Maniez was the French coordinator of ISKO in its early beginnings. In November 2000, the French Chapter adopted the non-profit legal organization status which enabled it to retain its independence from any organization in France and remains closely linked to International ISKO. This new status enabled the Chapter to draw conventions with various organizations and institutions where ISKO France could be housed so that they can take charge of specific events at the administrative and financial levels.

Since its creation in 1996 the French Chapter has organized, once every second year, micro conferences known as Journées ISKO-France. The first of these meetings took place in Lille in 1997 and, given their popularity, they were carried on since then and are now called, the ISKO-France Conference (ISKO-France website: http://www.isko-france.asso.fr/).

At the end of the Lille conference, the General Assembly was convened and a new Executive Board was elected: Prof. Dr. Amos David, University of Lorraine, was elected as the Chairperson, Prof. Dr. Widad Mustafa El Hadi as the Deputy-Chair, Prof. Dr. Gérard Régimbeau as Secretary and Dr. Philippe Kislin as treasurer.

The areas of interest of the French Chapter members and researchers can be seen through their contributions in the eight ISKO conferences held in France from their first meeting in 1997 to the last conference held in Lille in 2011 as follows:

- Epistemological and Historical approaches to KO
- Conceptual approaches to KO
– Abstract concept representation and automatic indexing
– Natural Language Processing techniques and their applications to cognitive analysis
– Semantic and discursive approaches
– User profile modeling
– Image indexing
– Indexing languages
– Monolingual and multilingual thesauri
– Statistical, mathematical & computational approaches applied to automatic indexing

– Design of information systems
– Social and Semantic

Reference

Knowledge Organization in Information and Communication Sciences, a French Exception?

Viviane Couzinet

Université Toulouse III Paul Sabatier, LERASS-MICS, 115 B, Route de Narbonne, BP 67 701, 31004 Toulouse cedex France, <viviane.couzinet@iut-tlse3.fr>

ABSTRACT: The alliance between information and communication sciences is a French specificity that originated in the 1970s from the necessity of assembling a sufficient number of researchers in order to obtain institutional recognition. The theme of knowledge organization brings a reflexive view on a discipline under construction. Our position in this article is to try, through a review of works conducted by the discipline's pioneers, to perceive how they envisioned the link between information and communication through the proposals made to their research community. French researchers approach the theme of knowledge organization in a way that does not seem very different from foreign research. As in foreign research, technique and technologies play significant roles. The ISKO conferences are, in this respect, very important. Knowledge organization also suffers from its interdisciplinarity, which deprives it of methodologies, theories, and concepts of its own. Its position at the heart of a discipline that is, itself, an interdiscipline, seems to authorize it not to consider its own fundamentals together with common theoretical foundations.

Received 7 March 2012; Accepted 7 March 2012

1.0 Introduction

It is often admitted that the alliance between information and communication sciences is a French specificity. This alliance would have originated in the 1970s from the necessity of assembling a sufficient number of researchers in order to obtain institutional recognition. If, in this respect, the scientific value of such an alliance could appear limited, the researchers considered as pioneers of this discipline, known first under the code “section 32” and currently under “section 71,” have been working on highlighting the strong links that unite these two disciplines.

The theme of knowledge organization, probably because it brings a reflexive view on a discipline under construction that was seeking academic recognition, was a main preoccupation of its founders for over ten years. Progressively, with technical matters prevailing over fundamental research and the institutionalization of the discipline considered acquired, French researchers were less bothered by their inscription in a field that had been identified as being an interdisciplinary. However, participation in a research field whose recognition might be constantly challenged by reorganizations imposed by the authorities relies on the collective elaboration of a common scientific project. Recent works have been following this process.

On the other hand, everyday it becomes less possible to consider science in relation to administrative or political frontiers that delimit its perimeter, methodologies, objects, concepts, and approaches. Thus, joining a position or founding an international trend en-
hances the debate and favors the progress of knowledge. Our position in this article is to try, through a review of works conducted by the discipline’s pioneers, to perceive how they envisioned the link between information and communication through the proposals made to their research community. Following the way this link is analyzed nowadays and recon-textualizing it in an international context, we will try to answer the following question: is there really a French exception in the way to conduct research on knowledge organization?

2.0 The elaboration of a link between information and communication sciences

2.1 Information and communication sciences

The definition of what the term “information and communication sciences” (ICS) covers has been the subject of the first publications of our discipline’s founders. Among them, Robert Escarpit and Jean Meyriat worked specifically on the way this discipline approaches its objects of study in order to distinguish it from other disciplines, the second focusing on matters that were then in the field of information sciences.

Having delivered, during the first Sofras’ congress in 1978, a definition of the notion of document that he would refine in 1981, Jean Meyriat associates information to the material object that supports it in order to communicate its substance and that cannot be separated from it. He opposes it to information processed by computer specialists, for whom information does not make sense and is limited to a combination of signals (Meyriat 1980). He also establishes a link with the notion of communication, a mental relation process used to pool it together with psychological, sociological, political, economic, and legal dimensions which explain why it is studied in all humanities and social sciences. What differentiates ICS from other sciences is that it involves the study of the communication process, its contents, and “the means it employs and the mechanisms that make it work” (Meyriat 1983, 82).

Communication implies assigning a meaning. It is generated through a set of elements. It proceeds from the relation established with other elements that generate a form with it. The role of the subject is perceived as essential, thus there is no information per se. The tight link between information and knowledge is perceptible through the shaping activity implied by the passing from one to the other. Information becomes knowledge when it is activated by the recipient in the interaction, who then integrates and assimilates it into his own stock of knowledge. This activation depends on its informative significance, on the individual or collective interpretation capacity, and on the situation in which it is taking place.

The connection between these two notions can be made clear. The term information refers to cognitive content and to “knowledge transmitted and acquired and that builds knowledge items.” The information that is the object of ICS work is then “knowledge either communicated or communicable” (Meyriat in Couzinet 2001, 251). It is then of “sustainable utility.” Jean Meyriat defines it as received knowledge which “is added to other knowledge that had been preserved and whose structured compound elements constitute knowledge that is enriched cumulatively” (Meyriat in Couzinet 2001, 151). It lends a capacity to act, and he describes it as “scientific” in the most general sense of the word. Its utility makes it essential and imposes its conservation, from which derives its very tight link with document and document memory, a secondary information device that implies a specific organization.

2.2 Knowledge organization: Jean Meyriat’s contribution

The first part of the research conducted by Meyriat on knowledge organization takes place in an international context. Involved since the beginning of the 1950s in a group of UNESCO experts, he set up a committee of social sciences experts. In 1980, he conducted a comparative analysis of over 50 information languages, defined as “linguistic tools used to describe specialized information and hence for analyzing and indexing documents, for storing and retrieving information, for building classified files and for operating documentation systems” (Meyriat 1980, 60). This research is part of a programme aimed at establishing a common language specific to these disciplines to ensure compatibility between indexing systems and thus improve international scientific cooperation in the field.

The corpus was made of languages that came from researchers’ projects or with a more operational status and an information retrieval function. These were either specialized or more general tools used to set up catalogues or thematic databases in the concerned disciplines. The first step in this research consisted of defining the fields covered by social sciences and then determining the number of descriptors that represented them in each language. The third step was a fine analysis of their semantic environment. It allowed
the author to highlight the elements that should be taken into account in the elaboration of what could be called an information language with a certain universal scope (Meyriat 1980), which would facilitate the exchange of research work.

Another matter of interest to the present conference is the epistemological work realized on the discipline through the organization of the objects covered by the field of ICS. Within a working group called “Communication means and contents” that he directed, Meyriat elaborated a classification of ICS. After renewed calls in various forms and articles submitted for review to members of the ICS scientific society or invitations to discussions at conferences—and that seem, as far as we know, to have received very few answers—the circumstances motivated him to establish the basis of a disciplinary territory. The Department of National Education had divided the discipline into five sub-domains in 1983—documentation, archives, information, communication, and community based activity—for which teaching and training should prepare. To this division based on professions, Meyriat proposes one based on epistemology. He thus defines four classes that include various branches of SIC.

In this classification, we can observe that the set of objects or sciences included in what is generally called information science is quite obviously seen in three subdivisions (communicology includes medialogy, which, in turn, includes bibliology, iconology, and documentology; informatology, communication technology, and social sciences of information clearly appear to belong to information sciences). However, a more careful reading of this classification allows us to perceive the interweaving between information and communication sciences. We can, in fact, easily consider, without seeming animated by any hegemonic intention, that some aspects of cinematology, mass communication, and functional communication also report to information science.

This work aimed at ordering the fields of our discipline shows the extent to which this order can sustain and even serve an epistemological reflection. The classification of ICS aims at building a common cultural context that expresses an identity capable of differentiating them from other academic disciplines. The preoccupation is info-communicational. It proposes an organized and visible content that aims at mobilizing and creating a feeling of belonging to a single family. This first proposal will be used by the National Council of Universities, the authority that manages the careers of French professors-researchers, to define the skill domains of this discipline. Established under the presidency of Meyriat in 1984 and transformed into a set of objectives and approaches, they have been regularly updated and are still in effect today.

2.3 Knowledge organization and scientific communication

There are then two ways of approaching knowledge organization. One is driven by the harmonization and building of a language, from content analysis and the determination of its structural characteristics. The object of the study is the meaning built by source languages in order to provide means to process information, to index it, and to make it retrievable. We could say that the operational scope of this work is the result of a will to contribute to the improvement of scientific communication. The other is based on an intention to make visible and to define a collective project included in a “territory.” Its political aim is to establish its positioning as a science. The object is not indexing anymore, but it is still scientific communication.

Manifesting an interest for “communicable or communicated knowledge,” “useful and sustainable,”
with the aim of facilitating its retrieval, but also its identification, because the role of the research subject is essential to have it make sense, implies defining and understanding its organization. This is one of the specialties of IS researchers. It has a historical background, a journal (Knowledge Organization), a scientific society (ISKO), and an undeniable international recognition within the field of IS. In France, it has also developed, and the organization of regular conferences by the French Chapter of ISKO gives visibility to this research area that has so little space in French journals.

We feel that to move forward on the proposal of establishing a tight link with scientific communication, we should aim at conducting works proposed by Yolla Polity in 1999 (Polity 1999) in a pluridisciplinary perspective. We must state here that in the recent past, at least three authors, who, for reasons of their own, did not institutionally join our discipline, E. de Grolier, J.-C. Gardin, and R. Pagès, have conducted projects that had a significant impact in the field. The last two have shed light on knowledge organization problems in the field of archeology and psychology. Polity proposes a chronology of French works: the 1970s, influenced by the expansion of document processing technology; the 1980s, oriented more towards natural language processing; and the 1990s, with the development of hypertext, interfaces, and neuronal networks. The recommended perspectives of “revisiting our cultural heritage in order to extract what topics remain and to build a shared conceptual body” and work on an “update of the domain modelization” (Polity 1999, 375) still seems relevant today.

We would like, eleven years later, to bring our contribution to this “state of affairs,” putting the emphasis not as much on the multidisciplinarity of the discipline, but rather on its intradisciplinarity. We agree with the idea that it is necessary to evaluate French scientific research because it appears, to us, in line with the wish expressed by Polity that this is a way “to allow the French research community to occupy a major position within the international research community” (Polity 1999, 375), and we also add the occupation of an original space that would ensure the link between information and communication.

Research that aims at highlighting roles other than those through which languages are generally analyzed is still little developed. The object is not to study the languages for themselves, but to transfer the focus towards their implication into a planned or induced communication. They then become material for observation. However, successfully conducting such investigations requires a good knowledge of these languages, of the way they work, are built, and used. We propose to illustrate the possibilities offered by this approach through three examples of projects conducted by our research team in Toulouse, some being centered on an institutional and epistemological pre-occupation, others on the construction of stereotypes, and then on informational culture.

3. Knowledge organization and communication processes

3.1 Institutionalization of a discipline

Starting with the idea that the institutionalization of a discipline goes through different phases and that it can be constantly challenged, a research project (Couzinet 2008) was conducted on a set of tools (classification schemes, bibliographies, databanks, book reeditions), elaborated in a professional field, that may contribute to the institutionalization process. The observation “ground” was that of information and communication sciences, and each version of the analyzed languages was anchored in a period going from 1970 to 1990.

The first phase is administrative. It refers to the agreement obtained at the highest state level, but it can only be consolidated if researchers make the works that they mean to develop visible. A second phase is then proposed that focuses on the interest of building a representation of the discipline. Languages are not, then, the only materials used. The approach is founded on the model developed by Estivals, who observed the evolution of bibliology through the tables of contents of books and classifications. Classification is considered by this author as a specific methodology of theoretical bibliology. It allows us to specify the object, the composition of the domain, and the theoretical perspectives under which phenomena can be studied. The classification scheme thus becomes a research scheme. Following this work, a classification of bibliography was elaborated by Estivals and Meyriat (Estivals 1993). It takes into account the French context and its inclusion into the field of ICS. It was followed by work that aimed at elaborating the Thesaurus of bibliography (Boustany and Estivals 1999), which made transversal links between sub domains and concepts visible. The two tools are intended to force delimitation and comparison with close disciplines and a reflection on the expression of concepts. Their elaboration leads to a representation of the domains that the discipline means to cover and of the space it occupies or means to occupy. This epistemological work corresponds to an internal construction phase.
The extension of this work (Couzinet 2009) consisted in proposing a third phase, one of consolidation based on the representation of the discipline in libraries. The observation focused on the space occupied within the classification and lists of subject headings used in these cultural places to perceive how the public is affected by them. The classification system of university libraries, within the framework of student training, was then analyzed. It shows, through the proximity between subjects, the development of the domain, including work on objects mainly within the field of communication sciences. Finally, the French newspaper Le Monde thesaurus defines a different vision, that of daily news and that shared with a larger public. It also highlights the representation of the work of documentalists by journalists, also considered information professionals, as being oriented exclusively towards practice.

3.2 The elaboration of stereotypes

To extend a set of research projects dedicated to the elaboration of stereotypes through concrete documentary objects, such as journals and magazines, Caroline Courbières proceeded to the analysis of the concept “feminine” through documentary languages. Seen as a “reference discourse in knowledge representation, while belonging to the cultural horizon through which they arose and on the contexts of reception in which they are interpreted” (Courbières 2010, 150), they are mobilized as texts that follow the elaboration of meaning in identified contexts. Claiming a documentological approach, the author needs to shed some light on their meaning and state why and how, the main hypothesis being that through their characteristics, they represent “a particular stereotype that fixes the feminine in its linguistic representations.”

The diachronic observation centers on abridged versions of classification schemes used during the second half of the 20th century (Dewey Decimal Classification, Universal Decimal Classification), and the UNESCO Thesaurus and RAMEAU authority file in their complete versions. The documentation centered on information processing is essentially semiological. It is then convenient for the analysis of the structure and concepts that make up its tools.

The documentary items show the image of a plural woman. Gender is the identity characteristic that includes the feminine concept in a second position through its sexual orientation. Sex as a practice is linked to the sphere of morality or law, but its value is reasserted in the mention of discrimination. However, the female sex is valued in the physiological domain through the notion of motherhood. The mother figure, in the social domain, refers to the single mother, the working mother, the housewife. Women’s work is associated to the figure of the citizen. Thus what Courbières (2010, 175 ff.) calls “documentary woman” is at the meeting point between transgression and domestication, and we cannot develop all its facets in this context. This documentary woman is represented in an asymmetry that is extended through the mentions of father and mother in the social sphere. “Oscillating between clinical discourse, community claims and singular social roles, the documentary language reveals fixed representations of the feminine role, at the crossroads between the private and the public spheres” (Courbières 201, 242).

3.3 Informational culture

In research conducted through a set of projects with the objective of making the importance of the link between professors and documentalists explicit, to point out the double responsibility assumed in French secondary school documentation centers, there has been an attempt to establish a link between the documentation activity and the teaching activity. A corpus has been gathered from a set of classifications and thesauri, privileging not the actuality of language, but rather the originality of presentation and its environment of application. Starting from the main mission of teaching, that is to help the student become an independent adult, the analysis was also centered on the thesaurus of Le Monde newspaper, taken as an opening on current affairs, and the thesaurus of the Chamber of Commerce and Industry of Paris, called Synchronized System of Economic Documentation (DES), taken as an opening on all possible professional activities.

The privileged entry point in these analyses is the distinction established on the basis of the training of documentation teachers between “information culture,” a generic name that we have proposed to use for the cultural fundamentals that each individual possesses or should possess, and “informational culture,” a culture that is specific to the community of information professionals. The latter is based on skills and knowledge acquired during courses referring to the ICS discipline. It is a capacity to mobilize practice and theory in order to ensure the transmission of intellectual methodologies and a knowledge authorizing the exploitation or appropriation of information with a distanciated and critical view (Couzinet 2008b).
The hierarchies and subject associations lead to an understanding of the meaning intended by the designer of the language. When a field of knowledge is not very familiar, it makes its approach simpler. However, the network of links is also an interpretation. A precise analysis of classification schemes underlines the fact that the main function of the tool is to manage a document collection, but it also has a communication function. This point of view, in the case of law, whose concepts are presented in the DES thesaurus so as to draw a balance, is particularly remarkable. Public and private law, which seem balanced, are situated on each side of an axis dominated by justice. If, in addition, we set business law as a foundation, the schema will carry a certain vision of Law. The use of colors, the arrangement of concepts, and the lines that link them mean that business surrounded by public law and private law may lead to justice (Couzinet 2011).

Using documentary languages to accomplish a technical task does not dissociate them from taking into account their operating and building modes, or from the context of the project underlying their elaboration. Training students to perceive the mediation that hybridizes itself through the project in order to facilitate information access, together with other projects made less perceptible by non-initiated, appears to us the main mission that justifies the union between documentation and teaching. In our opinion, it relies on the acquisition of an informational culture.

If we refer to the three major themes defined by Polity, these three examples show very uncommon ways, in France, to approach knowledge organization and its tools. They lie at the confluence of information and communication.

4. A French exception?

4.1. Foreign research

How is work on knowledge organization conducted abroad? We will not pretend to give an exhaustive answer to this question, which, on its own, would require extensive research. However, by referring to the synthesis proposed by Maria J. Lopez-Huertas, which is based on 151 references, and to articles from Birger Hjørland, we can perceive whether the approaches that we have isolated are present.

The most common research projects focus on the quality of knowledge organization systems, from the point of view of their content and also from the technical and technological standpoints. The key word in this way of tackling the subject is “interoperability.”

The tendency is to reformulate the questions in a technological and interdisciplinary context (Lopez-Huertas 2008), what Hjørland (2003, 88) considers the “technology-driven phases.” He isolates five of them: manual indexing and classification in libraries and benchmarking works that provide the principles of knowledge organization, which he considers still valid and important; documentation and scientific communication originating from the documentation movement founded by Otlet and Lafontaine; the recording and retrieval of information by computers since 1950; information retrieval through citations; and, finally, full text searching, hypertext, and internet searching since 1990.

This topic is not well supported theoretically and methodologically (Lopez-Huertas 2008; Hjørland 2003), it is rather made up of a superimposition of models and methodologies that are not really linked together. This is associated with interdisciplinarity, which is not specific to our discipline, but which reveals itself to be tricky, because it is, in fact, fragile. It is specified that, facing concepts imported from somewhere else, it is necessary to build our own terminology because epistemological problems severely affect the activity of interdisciplines (Lopez-Huertas 2008).

Another research orientation concerns social organization of knowledge. This trend is supported by Hjørland and Albrechtsen who consider it necessary to develop approaches based on a more historical and cultural understanding. This is what they call domain analysis. Involved in complexity, so they seem to refer to Edgar Morin (1990) and suggest domain studies should consider the complex interaction of ontological, epistemological and sociological factors influencing the development of fields of knowledge (Hjørland and Hartel 2003). The journal Knowledge Organization has published works that belong to this approach, in the artistic or nursing domains, for example.

4.2. French specificities

Is it possible to consider that there is, as the French like to believe in numerous areas, a French exception in the way we approach the matter of knowledge organization? If we compare the chronology made by Polity with the division made by Hjørland, it is possible to consider that these elements come up together, even if the aim of the argument of the first differs from the second, going into less detail and precision. The technological concern is omnipresent. Both call for the development of a theory.
The three research initiatives presented here could belong to the trend initiated by Hjørland, insofar as they take into account a social context and a knowledge domain and have a historical dimension. However, we also need to use our own knowledge of documentary languages, in the generic sense of the term, to highlight the processes that aim at influencing the users.

We consider that, by providing similar referents and ordering knowledge, languages produce and transmit a certain vision of the world. Within libraries, the ordering and presentation of documents, the signage, the address of each document recorded in the database, are repeated references facilitating the assimilation of an order. Some classification schemes are part of a project that surpasses the goal of document access. Gérard Régimbeau showed, for example, that in the artistic domain, document indexing contributed to the propagation of ideas (Régimbeau 1998). Similarly, classification schemes used in the USSR or China participated in the diffusion of a certain conception of collective life.

The power of suggestion of knowledge organization combines ideal elements that are often indispensable with practical information, in a constant interaction between a given situation and the individual. They induce a certain kind of behaviour, but also a way of thinking that provokes support for a project. As far as we know, the social approach promoted by Hjørland and Albrechtsen has not yet reached the communication sphere, but the first works in this trend tend to meet Meyriat’s and our own work.

4.3. About some exceptions

It is often admitted by French information and communication sciences researchers that the combination of these two disciplines—information science and communication science, and even other disciplines, such as cultural studies, media studies, museology—is an exception on the international scene. We agree with this, but we do not belong to those researchers who think this position has no scientific justification. The work conducted by discipline pioneers to precisely determine the limits of this exception clearly reveals that they can be imbricated. Specifying outlines is equal to defining research programmes.

If, historically, our discipline has left aside this fundamental exercise to ensure its perennity, studying its organization opens research perspectives at intersections. Thus knowledge organization can be thought of as (Hjørland 2008, 86):

Activities such as document description, indexing and classification performed in libraries, bibliographical databases, archives and other kinds of “memory institutions” by librarians, archivists, information specialists, subject specialists, as well as by computer algorithms and laymen. KO as a field of study is concerned with the nature and quality of such knowledge organizing processes (KOP) as well as the knowledge organizing systems (KOS) used to organize documents, document representations, works and concepts … In the broader meaning KO is about the social division of mental labor, i.e. the organization of universities and other institutions for research and higher education, the structure of disciplines and professions, the social organization of media, the production and dissemination of “knowledge” etc.

It can also be considered as the most convenient material to reveal social dimensions belonging to scientific, political, pedagogical or cultural communication. As such, it is not the privileged domain of information science. The complexity of its elaboration and its use as an indexing tool induces the capacity to analyze it as such. It is then one of the meeting points between information and communication, an intradisciplinary link.

Thus it is possible to question the plural form used in France, another exception, to name the discipline that studies it. Meyriat considers that the plural applies to press studies and the singular to documentation (Meyriat 1986). For Jacques Maniez, Information Sciences are “open to different aspects of this social phenomenon, whereas Information Science is wholly dedicated to theoretical and practical document matters, similarly to its English homologue” (Maniez 2002, 41). Multiplying research projects involving information and communication is part of the efforts made towards collective elaboration of the disciplinary project. The classification announced by J. Meyriat implies a reflection on an information-communication science, singular and autonomous, built around the link between significant and signified of an acquired and recognized maturity.

5.0 Conclusion

French researchers approach the theme of knowledge organization in a way that does not seem very different from foreign research. As in foreign research, technique and technologies play significant roles. Not
very welcomed in the French journals, and handicapped by language for its international circulation, its particularity is that it is not widely broadcasted. The ISKO conferences are, in this respect, very important.

Knowledge organization also suffers from its interdisciplinarity, which deprives it of methodologies, theories, and concepts of its own. Its position at the heart of a discipline that is, itself, an interdiscipline seems to authorize it not to consider its own fundamentals together with common theoretical foundations. The contribution of the ISKO journal, as a medium through which the international dimension of our research field is elaborated, is vital. However, it also is necessary to be concerned with the opening of French journals to this research topic in order to avoid isolation and to build the info-communicational approach of knowledge organization collectively.

Notes
1. SOFRASIC is the former name of the French Society for Information and Communication Sciences (SFSIC).
2. Eric de Grolier worked with J. Meyriat at UNESCO. Jean-Claude Gardin is the founder of the Synthol documentary language and Robert Pagès of the Codoc.
3. More recently, we can add René Ginouvès and Anne-Marie Guimier-Sorbets.

References

Metadata About What? Distinguishing Between Ontic, Epistemic, and Documental Dimensions in Knowledge Organization†

Claudio Gnoli

Università di Pavia, Biblioteca della Scienza e della Tecnica, Sezione di Matematica, via Ferrata 1, I-27100 Pavia, Italy, <gnoli@aib.it>


ABSTRACT: The spread of many new media and formats is changing the scenario faced by knowledge organizers: as printed monographs are not the only standard form of knowledge carrier anymore, the traditional kind of knowledge organization (KO) systems based on academic disciplines is put into question. A sounder foundation can be provided by an analysis of the different dimensions concurring to form the content of any knowledge item—what Brian Vickery described as the steps “from the world to the classifier.” The ultimate referents of documents are the phenomena of the real world, that can be ordered by ontology, the study of what exists. Phenomena coexist in subjects with the perspectives by which they are considered, pertaining to epistemology, and with the formal features of knowledge carriers, adding a further, pragmatic layer. All these dimensions can be accounted for in metadata, but are often done so in mixed ways, making indexes less rigorous and interoperable. For example, while facet analysis was originally developed for subject indexing, many “faceted” interfaces today mix subject facets with form facets, and schemes presented as “ontologies” for the “semantic Web” also code for non-semantic information. In bibliographic classifications, phenomena are often confused with the disciplines dealing with them, the latter being assumed to be the most useful starting point, for users will have either one or another perspective. A general citation order of dimensions—phenomena, perspective, carrier—is recommended, helping to concentrate most relevant information at the beginning of headings.

Adaptation of the paper for the proceedings of ISKO France conference held in Lille in 2011.

Received 7 March 2012; Accepted 7 March 2012

† Tom Pullman (Aberystwyth University and ILC project) provided precious terminological advice, in particular for the title and the names of dimensions; Thomas M. Dousa (University of Illinois) reported Otlet’s acknowledgment of the first two dimensions; Francesca Cattaneo (University of Pavia) provided plant images from Pavia Botanical Garden that were used in the presentation.
1.0 What is knowledge organization about?

For a long time, the most traditional form of indexing knowledge contents consisted of applying classification schemes and subject heading lists to printed books. However, new media have continuously appeared, the contents of which also needed to be organized: printed images, magnetic carriers, digital carriers, networked information, etc.

Beside this multiplication, we are now dealing with a convergence of media, through the universal language of digital formats, into integrated and diffused forms (cross-mediality): multimedia contents that easily pass from a mobile phone to a personal computer or a car navigator, interacting information devices in technologically equipped homes or retails, etc. (Resmini and Rosati 2008). The digital carriers are pushing libraries, archives, and museums to converge towards a common universal knowledge space (Rayward 1998), a trend confirmed by the increasing integration of cataloguing principles and schemes, such as FRBR or CIDOC-CRM, across library science, archive science, and museology. Knowledge organization (KO) is thus concerned not only with libraries, but with any collection of knowledge items including archived documents, natural specimens, and artifacts of any kind displayed in museums, galleries, and exhibitions, perhaps even organizations dealing with the subjects of interest (Gnoli 2010a; Latham 2012).

This situation poses new problems in identifying the boundaries of KO. If, for example, we state that KO deals with knowledge as recorded in documents, what should we consider as a document? The definition of notions like those of document, data, information, and knowledge is known to be difficult (Buckland 1997; Ridi 2010). Intuitively, we can say that a document is any carrier of information. However, as taught in semiotics, everything can convey information as it is interpreted as a sign of something other; the presence of a given plant can be interpreted as a sign that particular climatic conditions exist there which are necessary for the growth of that plant species. This would lead to the paradoxical conclusion that KO deals with everything.

Still, the domain can be restricted if we specify that conveyed information must have been intentionally put there to be interpreted by someone other. This rules out most plants, as they grow in a given place spontaneously, while only the plants intentionally put in a botanical garden to be displayed and illustrated by signs reporting their names are real documents. Which indeed makes botanical and zoological gardens, together with other kinds of exhibition, part of the scope of KO. In other words, as we are interested in subject contents, what matters is not the material object, but its use to convey knowledge.

2.0 The dimensions of knowledge organization

In 2007, I enjoyed the privilege of exchanging ideas about some general KO questions with Brian Vickery, an author whose work is recognized as central in the history of information science (Gnoli 2012). While discussing the role of disciplines and phenomena in classification, Vickery proposed this useful schema, later reported in a paper (Vickery 2010):

From the world to the classifier
– the world (nature, people, human artefacts) = phenomena
– people’s activities = disciplines, fields of activity
– reports of activity, each within the viewpoint of its own discipline (field)
– subjects of reports and of topics within them
– classification of subjects—which will need both disciplinary and phenomenal aspects

The schema makes clear how knowledge moves through a series of layers. The series originates in the real world, that pre-exists to knowledge and provides its objects. Real phenomena are studied by humans through their epistemic activities. These are structured according to various categories, including traditional disciplines. Documents can then be seen as reports about these epistemic activities, hence their content will include both structures of the original objects and structures of the activities by which they are investigated. Paul Otlet was a pioneer in acknowledging this more than one century ago, when he wrote that a classification “should enumerate both the objects and the points of view and choose as the basis of classification a sequence of one or the other as needs be” (Otlet 1990, 64).

To the features of the two previous layers, documents, in turn, add those of their own, like their format, length, or material. All these layers thus become part of the subjects that have to be identified and analyzed in classification (or, more in general, in KO). In other words, the reference of indexing terms and notations to reality is an indirect one through the mediation of documents (Hutchins 1979, 32-33).

I will call all these layers the dimensions of knowledge organization, following the use of this word by...
Tennis (2002) and Hjørland and Hartel (2003). Such a term expresses the fact that they are separate structures, all together concurring to form the subject of a document. The mathematical meaning of the term also suggests that the coexistence of several dimensions can be addressed by an analytic-synthetic approach, in which each knowledge item is ideally placed in a multi-dimensional space at the crossing of the coordinates for each dimension. Indeed, the analytic-synthetic model introduced in KO with facet analysis has been described as “multidimensional” (Gatto 2006). Notice, however, that in our model, facets themselves are to be identified only within each dimension: hence we will have the facets of phenomena, the facets of epistemic activities, etc.

Vickery’s scheme can be reformulated and extended in the following table, where each dimension is represented by a term, a corresponding field of study, and a symbol (a Greek letter, thus avoiding confusion with most KOS notations):

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Term</th>
<th>Field</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>(reality)</td>
<td>[mystique?]</td>
</tr>
<tr>
<td>β</td>
<td>phenomena</td>
<td>[ontology]</td>
</tr>
<tr>
<td>γ</td>
<td>perspectives</td>
<td>[epistemology]</td>
</tr>
<tr>
<td>δ</td>
<td>carriers</td>
<td>[bibliology]</td>
</tr>
<tr>
<td>ε</td>
<td>collections</td>
<td>[library science]</td>
</tr>
<tr>
<td>ζ</td>
<td>users</td>
<td>[sociology]</td>
</tr>
</tbody>
</table>

The next sections will consider the dimensions listed above in more depth, with special focus on dimensions β, γ, and δ.

3.0 The ontic dimension

Reality in itself (α)—what Kant called the *noumenon*—is perceived by humans only indirectly, through their sense organs and intellectual apparatus (with the possible exception of mystic knowledge, which we will not further discuss here).

Thus the actual basis on which KO can operate are the perceived phenomena (β): photons, granites, cats, teams, operas, etc. The term “phenomena” is adopted by various authors in KO literature (Mills and Broughton 1977, 49; Beghtol 1998; Szostak 2004, 30; Szostak 2007), although Dahlberg (2008) finds it misleading and prefers “general objects.” The identification and ordering of phenomena is the task of ontology, the study of what exists, now increasingly applied to the organization of digital knowledge. Phenomena are often opposed to the disciplines studying them, as an alternative starting point for the organization of knowledge, especially in general classifications (Mills and Broughton 1977, 55): we can choose whether to first consider the phenomenon “stars” or the discipline “Arabian astrology” that studies it under a particular perspective.

Many disciplines can be described as the scientific study of a given class of phenomena, like astronomy is the study of stars, botany is the study of plants, etc. However, for Mills and Broughton, these are only “sub-disciplines” of a smaller number of “fundamental disciplines,” like science, philosophy, history, and art, which can be defined in epistemic terms, as alternative “ways of looking at the phenomena of the world;” history could then study everything in a chronological perspective, art could represent everything in creative forms, etc.

While disciplines are traditionally adopted for the organization of printed books, it can be difficult to apply them to the greater variety of contemporary media. In this sense, phenomena offer a more generalizable basis that can be shared between very different media (Gnoli 2010a), because, as is shown in our scheme, they are a more fundamental dimension of knowledge: an Arabic parchment, a documentary film, and a planetarium presentation can all refer ultimately to “stars.” In the words of librarian Douglas Foskett (1970, 45): “reality is the basis for the texts of documents; that is what authors try to describe, and what searchers are investigating.” More recently, philosopher and computer scientist Barry Smith stated similarly that ontologies are concerned with “building models of entities in reality, thus for example building models of the organization of the genome and not just of information contained in this or that database” (Smith 2004, 77 emphasis his).

Of course, the ways in which reality is analyzed into distinct concepts depend on the current advancement of knowledge; concepts like “aether” or “phlogiston,” although originally intended to denote real phenomena, have subsequently been found to be inappropriate and abandoned, while other concepts have changed in meaning as knowledge progressed (LaPorte 2004). The consequence of this for KO is that KOSs will always need to be updated. Even the ontic dimension of knowledge depends both on reality and on theories about it (Popper 1972). The extent at which theories determine concepts is widely debated in philosophy. Still, given a certain stage of development in knowledge, phenomena can be conceived as entities separate from the ways to study them.
4.0 The epistemic dimension

Phenomena coexist in subjects with the material and intellectual means by which they are considered: microscopy techniques, semiotics, Marxism, poetry, education of children, etc. These include the disciplines, as discussed above, but also the domains addressed by different research communities (Hjørland 1995), the human activities to which knowledge is intended to be applied (Vickery 2008), the communicative functions performed in transmitting knowledge (Hutchins 1976, 8), the theories adopted and methods applied (Szostak 2007), the historical epoch and geographical context in which knowledge is produced (Tennis 2002), and, in general, all viewpoints adopted by authors (Beghtol 2002).

In our scheme, we have subsumed all these under the label of perspective; this term, as well as others like aspect, viewpoint, or bias, have been used to describe KOSs that organize not phenomena directly, but rather ways of looking at them (Langridge 1992, 6-10; Svenonius 1997; 2000; Slavić 2007). Perspectives can be studied by epistemology, the science of the ways and means by which knowledge is acquired.

A faceted classification able to distinguish between different knowledge dimensions, like the Integrative Levels Classification (ILC) (Gnoli et al. 2008), may represent all the kinds of approaches mentioned above as facets of the epistemic dimension, as opposed to facets of the ontic dimension. In ILC, facets of the epistemic dimension begins by the digit 0 and are listed in the following table:

<table>
<thead>
<tr>
<th>0</th>
<th>perspective</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>epoch</td>
</tr>
<tr>
<td>02</td>
<td>place</td>
</tr>
<tr>
<td>03</td>
<td>method</td>
</tr>
<tr>
<td>04</td>
<td>theory</td>
</tr>
<tr>
<td>05</td>
<td>discipline</td>
</tr>
<tr>
<td>06</td>
<td>culture</td>
</tr>
<tr>
<td>07</td>
<td>activity field, domain</td>
</tr>
<tr>
<td>08</td>
<td>modality</td>
</tr>
<tr>
<td>09</td>
<td>communicative function</td>
</tr>
</tbody>
</table>

It can be noticed that, while in many documents, phenomena are the primary object of treatment, particular kinds of documents exist in which phenomena are less important as compared to perspective: that is, it is not very important what is represented, but how it is represented. Examples of this are poetry and other forms of art, in which very different objects can be represented to convey one same message, like sadness or nationalism, and political cartoons, where the represented phenomena are often allegories expressing a political judgment rather than the actual details shown (Landbeck 2008).

5.0 The documental dimension

A third relevant dimension in our present analysis is given by the formal features of knowledge carriers: videos, MPEG formats, dates of production, durations, document sizes, etc. These contribute the subject matter of documents with a layer δ that, although less basic than those of phenomena and perspectives, may nevertheless get some relevance.

ILC analyzes this dimension into the following documental facets, beginning with 00:

<table>
<thead>
<tr>
<th>00</th>
<th>document</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>publication time</td>
</tr>
<tr>
<td>002</td>
<td>publication place</td>
</tr>
<tr>
<td>003</td>
<td>language</td>
</tr>
<tr>
<td>004</td>
<td>medium</td>
</tr>
<tr>
<td>005</td>
<td>section</td>
</tr>
<tr>
<td>006</td>
<td>author</td>
</tr>
<tr>
<td>007</td>
<td>target</td>
</tr>
<tr>
<td>008</td>
<td>commented document</td>
</tr>
<tr>
<td>009</td>
<td>format</td>
</tr>
</tbody>
</table>

Like with perspectives, carriers also get special importance in some kinds of documents that are strongly formal. This is the case with abstract paintings or instrumental music, which can hardly been said to represent any specific phenomenon. Exceptions are still possible, like Bedřich Smetana’s The Moldau, an instrumental symphonic poem that explicitly refers to an actual river (phenomenon), with music imitating the flow of the river in its various stretches, and more implicitly to the ideal of Bohemian national identity (perspective).

Further, pragmatic layers concerning the storing and circulation of knowledge contents can be identified, like those of the particular collection in which a document is kept together with others, or the particular community of users that interact with it. However, we will not consider these dimensions in detail here.

6.0 Representing the dimensions

The three dimensions that we have discussed in detail manifest themselves in actual documents in various ways. Ranganathan wrote that a book has a mind (the phenomena it deals with), a language (the perspective
adopted in doing so), and a body (its material carrier) (Ranganathan 1967). In metadata, dimensions can be combined more or less explicitly; consider such titles as “Handbook [carrier] of African anthropology [phenomena + perspective],” “Lectures [carrier] on set [phenomena] theory [perspective],” “Bird [phenomenon]-watching [perspective] in the Cotentin peninsula [phenomenon]: a guide [carrier].”

In informal communication, like e-mail subjects or webpage titles, carriers and perspectives are often provided without reflection as the first or even the only knowledge element: “Information on ...,” “Question.” Clearly, such metadata are much less useful than if phenomena were given precedence and used as main labels.

The latter strategy would correspond more closely to what is taught in many handbooks of subject indexing, which recommend to leave formal specifications, such as “guide,” at the end of compound strings. A similar principle is used in classified shelfmarks, where metadata belonging to the documental dimension, such as date of publication or initials of the first author, are expressed (if at all) only after the symbols for the basic subject content (perspective + phenomena). In many cases, indeed, the most relevant information—also called the main theme in subject indexing (Buizza 2011; Gnoli 2010b)—is what a document is about, while its approach and form are only complementary specifications.

It is not by chance that digital interfaces using resizable windows, like Web browsers, when fed with a string of metadata longer than the available space, are programmed to display its beginning rather than its end. Therefore, for the purposes of information architecture, a principle of front loading has to be recommended, consisting in concentrating the most relevant information towards the beginning of a string.

In general, a recommendable standard citation order between dimensions is:

phenomena > perspective > carrier

As we have seen, classical bibliographic classifications reverse the first two dimensions by taking disciplines as their first divisions. This is, in itself, a perfectly legitimate alternative, whose efficiency could be tested and compared. Comparison would need that the distinction between phenomena and perspectives were clear, as is also recommended by Svenonius (1997, 16). However, disciplinary classifications can mix these two dimensions in shaded ways. UDC class 59 codes for the discipline “zoology,” while its subclasses have captions with nouns of phenomena, like 599 “mammalia, mammals.” In the faceted perspective now adopted in UDC, subclasses like “mammals” are interpreted as the first facet of zoology, belonging to the general category of Things, although not separated from the discipline class in the notational plane. Distinction between discipline and phenomenon can result in benefits for machine treatment.

Confusion between dimensions can be observed in many information resources and tools. The application of facet analysis to Web information architecture has enjoyed much success in last years (La Barre 2004), having recently been adopted even in Google search results. However, what information architects call “facets” are often facets of the documental dimension, such as date, size, or language, which are easier to obtain and to treat automatically, while the original notion of facet was developed in library classification with reference to the more substantive facets of the ontic and epistemic dimensions, such as part, process, or agent.

This confusion seems to be spreading in metadata terminology too. The development of ontologies and the very notion of a semantic Web have arisen just in response to the lack of tools to organize and connect digital contents by their subject matter, while tools for managing descriptive metadata—such as the Dublin Core elements set—already existed. However, the success of the new tools is now reflected in calling “semantic” even metadata for descriptive indexing, including “ontologies” for description of documents by authors, title, date, etc. Again, it seems that a clearer distinction between the dimensions identified in this paper will be increasingly useful.

To summarize, our general thesis is that there is a need for distinguishing between the different dimensions of knowledge items and for treating each dimension separately in an appropriate way. This thesis agrees with the five recommendations of the León Manifesto (ISKO Italia 2007):

The current trend towards an increasing interdisciplinarity of knowledge calls for essentially new KOSs ...
this innovation is ... feasible
instead of disciplines, the basic units of the new KOS should be phenomena ...
the new KOS should allow users to shift from one perspective or viewpoint to another ...
the connections ... can be expressed and managed by analytico-synthetic techniques.
These requirements are being implemented in the experimental ILC system. As reported above, all facets conveying information on perspective and on carrier, as opposed to phenomena, can be identified in ILC by their facet indicators. This allows for parsing phenomena, perspectives, and carriers as separate dimensions in compound classmarks, and for their automatic treatment in digital applications—e.g., displaying each dimension in a different font, displaying only some dimensions, displaying dimensions in alternative citation orders, search and extract only items with a given phenomenon, or perspective, or carrier independently from the other dimensions.

ILC perspective facets are especially tested in the BioAcoustic Reference Database, a classified bibliography where facets of scientific method are often relevant (e.g., “harbour porpoises, nervous system, studied by magnetic resonance”) (Gnoli et al. 2008).

7.0 Concluding remarks

Traditional KOSs that mix more than one dimension into simpler headings, like disciplinary classifications, do so under the assumption of literary warrant: if documents have been produced by their authors with some perspective and form, they will be useful to users adopting the same perspective and working with the same forms—say, only users working in the domain of zoology or only users working with online resources. This approach reflects a conception of the task of KO as limited to the representation of available sources in a faithful way. It tends to produce conservative applications: research communities will continue to read and cite only themselves, without taking advantage of what has been done by applying other perspectives or other carriers to the same phenomena, or by considering other phenomena by the same perspective, etc. (Szostak 2007).

However, one can also conceive that KO do more than just keeping the status-quo of knowledge; it could also highlight previously unnoticed connections between existing knowledge that will stimulate further research (Davies 1989). In order to enable the creation of new knowledge across different domains, disciplinary schemes should be replaced by more flexible structures (Jacob 1994).

This seems to be possible only if the different dimensions of subject matters are analyzed, identified, and represented separately so that each one can be searched and retrieved alone and creatively associated with others. While perspectives and carriers can provide important specifications and sometimes even become the main theme, the most universal knowledge units, on which an analytico-synthetic KOS should be based, are phenomena.

References


Szostak, Rick. 2007. Interdisciplinarity and the classification of scholarly documents by phenomena, theories, and methods. In Rodríguez Bravo, Blanca, and Alvite Díez, María Luisa eds. Interdisciplinarity and transdisciplinarity in the organization of scientific knowledge: actas VIII congreso ISKO-


The Catalog Resisting the Web: 
An Historical Perspective

Philippe Bourdenet

SCD de l’Université du Maine, Av. Olivier Messiaen, 72085 LE MANS Cedex 9, France, <Philippe.Bourdenet@univ-lemans.fr>

ABSTRACT: Libraries are currently seeking to restructure their services and develop new cataloguing standards to position themselves on the web, which has become the main source of information and documents. The current upheaval within the profession is accompanied by the belief that libraries have a major role to play in identifying and supplying content due to their extensive high quality databases, which remain untapped despite efforts to increase catalog performance. They continue to rely on a strategy that has been proven successful since the mid-nineteenth century while seeking other models for their data. Today, they aim to exploit changes brought about by the web to improve content identification. The current intense debate on RDA implementation mirrors this desire for change. The debate is rooted in past efforts and yet tries to incite radical changes as it provides for interoperability from the creation of records through an object modeling in line with web standards and innovations. These innovations are presented through an historical perspective inspired by writings by librarians who are entrusted with helping in the development of bibliographic description standards.

1.0 Introduction

Standardization activity in libraries has probably never been so intense nor raised as much concern (Danskin 2006) as it currently does with the debate surrounding the implementation of RDA in France. In order to better understand the challenges that standardization presents, one must consider historical context and the underlying problems not necessarily linked to RDA itself but to the model upon which it is based, FRBR, whose emergence demonstrated an extensive review on how to improve information cataloguing and identification tools.

For libraries, the problem is twofold, because RDA clearly intends to meet the users’ needs as interpreted by RDA developers, who may have a distorted understanding of some necessities or priorities. It supposes that users’ needs will be met by an overhaul of cataloguing rules. This very hypothesis may be biased. It is indeed an approach that has been proven effective in the past. Over the past 150 years, libraries’ cataloguing solutions were based on this hypothesis and were successful until the end of the 20th century. But it is unclear whether applying this principle today would bring similar success. Until the 1990s, little progress had been made in the container/contents relationship, but problems of management and identification have dramatically increased in direct proportion with the explosion of digital material in its native form as well as user accessibility and tools. FRBR is driven by the necessity to create content metadata whereas cataloguing rules, in their most evolved form,
are still producing data based on containers (Moore 2006).

2.0 Innovations in libraries

2.1 Historical Heritage

FRBR evolved out of a series of developments in the history of cataloguing (Taylor 2007). The famous 91 rules of cataloguing established by Sir Anthony Panizzi (1841), which aimed at improving the quality of the descriptions of the British Library printed books, were already highly innovative. Another development worth mentioning is Charles Cutter’s cataloguing code, known for the Cutter number which integrates an abbreviation of the author’s name in the classification code, as well as the third edition of its manual Rules for a Dictionary Catalogue (1904), which elaborates on principles first presented in 1876 and gradually enhanced. His definition of a catalog addresses three objectives: to “enable a person to find a book of which either the author (A), the title (B), or the subject (C) is known, to show what a library has by a given author (D), on a given subject (E) or in a given kind of literature (F), and to assist in the choice of a book as to its edition (G) (bibliographically) or as to its character (H) (literary or topical).” These briefly summarized basic rules require the creation of an “author-entry (for A and D), a title-entry (for B), a subject-entry, cross-references and classed subject-table (for C and E), a form-entry and a language-entry (for F), giving edition and imprint, with notes when necessary (for G) and notes (for H),” in short, nearly all the information contained in a record, even today. The system does not propose a solution for identification of multiple authors of collective publications, pseudonyms, or rules for author classification, collective responsibility or indexing resources with non-Latin alphabets. Despite these shortcomings, which must be addressed especially since the volume of collections is growing, FRBR remains the distant successor to the rules proposed by Cutter. The connection is obvious between the statements “find a book of which the author is known” and “find a particular manifestation when the name(s) of the person(s) and/or corporate body(ies) responsible for the work(s) embodied in the manifestation is(are) known” (Taylor 2007) (FRBR 1998).

As early as 1900, the United States began to harmonize cataloguing rules from the ALA, the Library of Congress, and the Dewey Decimal system with cooperation back and forth across the Atlantic, which took into account new British cataloguing rules and “Prussian Instructions” used in Germany and some Scandinavian countries. The international cooperation, in its early stages, was supported by the work of the mathematician and librarian Ranganathan who, in 1931, wrote the Five Laws of Library Science:

Books are for use
Every reader his (or her) book
Every book its reader
Save the time of the reader
The library is a growing organism.

The word “growing” in this context does not only refer to libraries’ collections in terms of linear meters of occupied shelves, but also to expansions in technology, knowledge, and international flow of knowledge, ease of global transport and migration of populations especially those in charge of their culture, conflicts, etc. of which the library is a reflection. In addition, these evolutions introduce the concept of use, a preoccupation addressed by FRBR (Taylor 2007)—inherited from AACR—and the overhaul of cataloguing and indexing rules carried out by IFLA for more than 40 years now and initiated by the Paris principles (1961), which put user needs at the centre of the debate. Changes in the uses and mentalities which allow us to perceive the library as an organization have taken a major step forward with digital technology because information is more easily manipulated by the user, allowing an intangible appropriation of the information. This appropriation is only possible if some of the mechanisms of data processing are made transparent. This need “puts the library community in the position to move forward quickly into the broader world of information exchange and reuse outside the library silo created over the past 40 years” (Hillman et al. 2010).

2.2 The catalog as a factor of opacity

Before considering data recovery and manipulation, which are relatively new needs, we must focus on the catalog’s main functions. Its “first objective is to enable the user of the catalog to determine readily whether or not the library has the book he wants” (Taylor 2007). This going back to basics may seem extreme according to Lubetzky (Taylor 2007), who, in 1953, considered certainties which, over the years, may have been overlooked by bibliographic search innovators. An example of one such certainty is that libraries consider their catalog a tool that renders
transparent their primary activity of building up collections while the user (a layperson) is unable to maximize use of the catalogue without specialized knowledge and/or training. The second main objective, subtler but linked with the first one is “to reveal to the user of the catalog, under one form of the author’s name, what works the library has by a given author and what editions or translations of a given work” (Taylor 2007). Even if it appears as a revisited rule, the concept of “work” is finally introduced here, for the sake of identification, and shows indirectly one inadequacy of Cutter’s vision whose searching principles were focused only on books.

Since 1961, initiatives to develop description models and cataloguing tools multiplied; in 1967, (Manning 1998) AACR’s first conference already suggested new codes, in 1990 (FRBR 1998) in Stockholm, economic realities were first discussed including the pressure on institutions to reduce cataloguing costs while improving user services and providing tools of description for different media.

2.3 The FRBR model

FRBR (Functional Requirements for Bibliographic Records) was first issued in 1998, after ten years of work. It is a conceptual model, which, if implemented, provides the user with a catalog allowing him to find, to identify, to select and to obtain a resource [IFL 01]. The FRBR can be briefly presented here bearing in mind that its specifications are based on a model of 10 entity-relationship, sorted into three groups which cover respectively (1) the level of bibliographic description (work, expression, manifestation, item), (2) the level of responsibility (person, corporate body) and (3) the level of subjects of works (concept, object, event, place).

In 1997, faced with inevitable dissatisfaction of both professionals and users due to inadequate cataloguing rules for the new media while the AACR were being reviewed, the Joint Steering Committee (JSC) held a conference focused on future uses of this cataloguing code and invited international participants “in the hope of freeing cataloguing rules establishment from the Anglo-American context only and developing a code which could be used worldwide” (Taylor 2007).

RDA is a cataloguing code which is based on the traditional principles of Panizzi, Cutter, Lubetzky and which also validates the principles of the FRBR conceptual model and its extensions in the field of authority control (FRAD, FRSAR). The seven chapters on description summarize FRBR intentions in addition to the first one which is a general rule: (2) Identification of the resource (see “identify”); (3) Type of media (see “select”); (4) Content (see “select”); (5) Access (see “obtain”); (6) Persons, families, and corporate bodies; and (7) Linked resources (see “find”).

RDA therefore proposes a set of instructions allowing the creation of metadata to describe a content on a standardized form for the Web, able to take account of either an electronic resource or a paper document. Some of the leading bibliographic agencies are currently examining implementation options for their bibliographic models to decide whether or not to adopt RDA. Though this can have immediate benefits for bibliographies (Pisanski, Zumer, and Aalberg 2009), the objective is to improve catalogs, with their bibliographic records (B) and their authority records (A).

3.0 Implementing FRBR in RDA

3.1 Implementation scenarios worldwide and in France

Economic stakes are very high. The AACR had already intended to provide solutions to reduce cataloguing costs, by facilitating data exchange, harmonizing practices in the Anglo-Saxon world and supporting derivative cataloguing. The FRBR model and its implementation in RDA require so many changes both in the presentation of data and in their intellectual orientation through introduction of the object model in order to progress towards the Semantic Web vision (Dunsire 2009) that original MARC records are unusable. The Library of Congress currently estimates the number of MARC records at just over one billion (McCallum 2004) worldwide, as a result of collaborative work over the past four decades. Tom Delsey, RDA Editor and member of JSC, defines three implementation scenarios for the databases currently managing MARC formats (Delsey 2009):

-First scenario: “RDA data are stored in a relational or object-oriented database structure that mirrors the FRBR and FRAD conceptual models” (Delsey 2009). These structures are not those of current MARC formats.

-Second scenario: “RDA data is stored in database structures conventionally used in library applications [MARC]. In those structures, data is stored in bibliographic records and in authority records, and in some implementations in holdings records as well” (Delsey 2009). Data would be modified in order to establish bibliographic descriptions or authorized access points.
representing FRBR entities, and bibliographic records will be linked to authority records for persons or corporate bodies, expression, manifestation, or work.

-Third scenario: RDA data is stored in bibliographic and authority records, based on the entities model, with no link between them.

In France, cataloguing standards (CG 46 "Information and Documentation," 28 March 2000) are prepared and updated by the AFNOR Group CN 357, which appointed the GE6 (Expert panel number 6) to do a study on RDA implementation, the feasibility of the above scenarios, the associated development costs for document applications, and to provide a solution to ensure data recovery by the system.1

There is no easy way to “FRBRise” a MARC record. The first scenario, ideal because it complies with FRBR requirements by creating records which can be exploited for web application, would be extremely expensive to implement. It would be contradictory for the AACR to adopt this solution since their objective is to streamline expenditures. It wouldn’t be easy to take this decision—even if it provides “an interesting theory about a four-level model—especially that the FRBR have not been tested in actual practice” (Coyle 2004). Librarians therefore don’t have the necessary hindsight or enough practical experience to refute or support it (Pisanski, Zumer, and Aalberg 2009) (Maxwell 2008), despite the new initiatives for its application with varying degrees of rigor in the United States, Australia, Sweden and even in France.

The second scenario would be more reasonable cost-wise, but it would require creation of FRBR entities from existing MARC records. This tedious work on evolutions based on strict mappings would obviously take time, but would thereby allow a gradual test of the improvements to the information systems, provided that user response would be able to be gauged along the way through accessible functionalities (Wells 2007). This expertise largely lies with bibliographic agencies and providers of MARC records in France, mainly the BnF3 and the ABES,4 and the decision is up to the strategic committee of CG46.

3.2 Tradition and innovation

The profession will most likely face some major upheaval, but experience has shown that, with regard to libraries, change cannot not be rapid, because even very dynamic evolutions must be rooted in a principle of sustainability due to the stability of the profession, its uses, as well as its trade practices (Calhoun 2007), but proposed JSC deadlines are too short. Some accomplishments show that libraries are willing to change their practices and adopt other non-traditional formats, as was the case at the University of Arizona, in cooperation with the NAL,5 with the creation of a digital library in RDF (Han 2006). The initiative of creating web standards for library use with the W3C Library Linked Data Incubator Group also provides evidence of what librarians are ready to do to position themselves on the web, which has become the main source of information. The adoption of object-oriented models able to make bibliographic information at the level of the Semantic Web is inevitable and the theoretical model developed by FRBR is, from this point of view, of enormous value, but will RDA allow this? Probably disappointed by JSC decision to convert AACR2 from RDA—a courageous but risky choice—librarians have shown strong resistance to the adoption of cataloguing code in the United States. The decision raised reactions as well as criticism from librarians over RDA, because “it neither sticks with the standards we’ve already got, nor offers anything [the] present OPACs can make use of in any kind of a helpful way”6. This code is still seen as a “prediction,” which “could theoretically work in the future,” but which has a long way to go to prove its value. Librarians’ distrust of RDA is joined by a more scientific criticism that questions whether libraries’ evolution should include RDA implementation. It is perhaps a mistake to start with cataloguing rules development to change the services (Coyle and Hillmann 2007): “Prior to elaborating detailed cataloging rules for libraries, we need to decide whether the user will view a general bibliographic tool that connects users and information resources no matter their origin, or continue to view a library inventory, that requires users to look elsewhere for other information they might need.” This is an uncomfortable issue for the profession, as it highlights the risk in introducing profound changes that may cause years of inaction before providing theoretically better service. There’s currently no proof of RDA’s effectiveness, which is all the more risky given the urgent need to rethink the services, to give them meaning and to make them compatible with users’ practices. This move could turn out to be a waste of time and an inappropriate investment—to make a fundamental change in services—when libraries urgently need to reach new users, build stable foundations when, at least in the near future, they must continue to offer user tools universally regarded as obsolete, even within the profession, with the un-
certain prospect of creating more appropriate services. There is the risk that readers may permanently turn their backs on inaccessible document platforms, including the traditional OPAC.

3.3 Fears and suspicions

Reservations expressed in North America may be explained by the fact that, unlike UNIMARC, MARC21 does not manage the links between the records with the identifier indexing, which dramatically increases the adjustments to move towards RDA, but this is only one factor. An overview of these reservations reveals suspicions about the real objectives, sometimes by the strongest supporters of the FRBR model (Coyle 2004; Coyle and Hillmann 2007). Yet, there has been extensive communication efforts on RDA objectives at the very core of RDA toolkit specifications. “The first objective of RDA is to be sensitive to the user needs” (Oliver 2010). Each chapter of the RDA toolkit indicates the functional objectives and principles (see for example “Record Attributes of... / Section 1 : Manifestation & Item / ... / 1.2 Functional Objectives and Principles”).

4.0 Challenges and perspectives

4.1 A great desire for change

The objectives as presented in the RDA toolkit are highly persuasive since they are based on practical considerations. If we are to still speak of “records” as aggregate data forming an intellectual unit to describe a resource, the resource is an entry to a set of services which are more or less extended according to the user authentication and then more or less personalized. More generally, a search in a FRBRised catalog should allow the disambiguation of a result, mainly in view of improving navigation and display of information. The proposed granularity is finer than it is for MARC or ISBD records. Catalogers are no longer allowed to note information that serve to identify a resource through a long character string which concatenates disparate properties (for example, the UNIMARC General Note block 3XX). Instead, all information is qualified. A record aggregates metadata on information that is an integral part of the resource by describing its properties through a pattern, the projection speed of moving images, the device or software required to read a document on a medium, or any other associated material, rather than the “Other material characteristics,” which is a very vague and ambiguous indication found in the full display records generated by traditional OPACs. In addition, RDA offers removal of abbreviations, indexing of the whole statements of responsibility which so far were limited to three, etc.

4.2 User benefits

The Results presentation allows a classification of resources according to their nature (with metadata focusing on content). A search for “le Barbier de Séville” (The Barber of Seville) using a catalog can return as diverse results as theater plays, prints, critical texts, music scores, video clips, DVD-ROM references, etc. (many thousands with the BnF’s general catalogue)—with specific icons for each medium. Instead, RDA would propose a quick answer with potential analysis of series of results listed by content type and this would be the first visible effect of a modeling work. This achievement would also allow us to widen search for particular adaptations, parodies, or other works which are intellectually linked to the original search and identify the resource main language from the short list of results. The manifestation of these aggregates requires complex data management to create links between the data. Of course, this would be possible with traditional tools, but engineering efforts would be of such a magnitude that barriers would seem insurmountable. The RDA model creates functionalities by using information existing in the data and no longer trapped in relational databases. The traditional catalog is seen as a silo whereas RDA introduces a paradigm from the creation of records, which do not need to be stored in complex and invisible reservoirs. And this, on its own, facilitates the readability of web holdings and communication with other communities producing data, because it is not designed for a particular format. The obvious advantage is that users can combine searches from more than one source, (by) bringing together users of ONIX’ (Kiorgaard 2006), which is a standard designed for the industry, and academics who use bibliographic tools complying with the standards applied in libraries. Web standards allow this convergence and RDA, based on these technical prerequisites, offers the hope of common semantics and data reuse in other contexts.

4.3 Institutions benefits

From an institutional perspective, to seize this opportunity is to capitalize on technological innovations
from other countries for common implementation the boundaries of which are defined by the web technologies. A recovered record inherits all the links it contains and introduces the concept of “continuity.” Acquiring a record is recovering more than one link in the chain. Since it can be reused by another institution or a user, this link will allow the extension of another chain and connect two worlds (RDA Toolkit, Key Features, 0.1). Though FRBR is intellectually a successful model, it is sterile, with no mechanism or tool to create data. RDA is this tool and, in addition, proposes a set of benchmarks to help make the right decisions regarding hierarchical arrangement of information and strategic orientation (in terms of services strategy) to reach a category of users. RDA supporters as well as the undecided are on the lookout for arguments the likes of which they usually obtain with technical case studies, but information and documentation professionals have to agree on the need to pursue this objective: “bibliographic data were created to be read and understood by librarians and users” (Coyle 2007) at a time when OPACs have been criticized for creating misunderstanding between librarians and users.

5.0 Conclusion

These considerations make a critical viewpoint of RDA more difficult, since libraries seem determined to evolve towards the Semantic Web. There aren’t enough alternative initiatives for this evolution to categorically reject the RDA model and production tool. This work requires consideration of RDF schema: "labels, areas and attributes need to be expressed as classes and properties" (Dunsire 2009), which is the basis of the object orientation, and vocabularies will have to evolve into SKOS and semantic relations into OWL. RDA proposes to develop straightaway interoperability of data from its creation, whereas the work done in the development of document platforms over the last decade implemented interoperability protocols either through connectors to build bridges toward external reservoirs or through local integration of data created elsewhere, which now seems to be inadequate to exploit the data from the web. This task was possible thanks to the high quality control on data, and this kind of exchange requires stability and continuity. Integrating the need for interoperability from the very outset stems both from tradition and from a very challenging dynamism—tradition because strictly respecting data formats is not new and ensures readability by other institutions which will now be able to understand and use them, and dynamism, because the method uses processes of relations with entities, creating dynamic interoperability with semantics which allow them to be used on the web and not only in libraries. Overall, this is not fundamentally different from the changes proposed by Panizzi, Cutter, Raniganathan, or Lubetzky, who saw a need to work on records’ substance in order to improve services and take into account global technology advances to meet user needs.

Notes

1. RDA Toolkit, Introduction, 0.2 : Relationship to Other Standards for Resource Description and Access
3. Bibliothèque Nationale de France (= French national Library)
4. Agence Bibliographique de l’Enseignement Supérieur (= French Higher Education Bibliographic Agency)
5. National Agriculture Library

References


Moore, Julie Renee. 2006. RDA: new cataloging rules, coming soon to a library near you! Library hi tech news 23n9: 12-17.


The Role of Activities Awareness in Faceted Classification Development

Orélie Desfriches Doria

Laboratoire DICEN, Case courrier 1D6R10, Conservatoire National des Arts et Métiers, 2, rue Conté, 75003 Paris, <orelie.desfriches_doria@cnam.fr>

ABSTRACT: In this paper, we propose a part of the methodological work to accompanying the development of a new type of Knowledge Organization System (KOS) based on faceted classification. Our approach to faceted classification differs from its traditional use. We develop a theoretical typology of professional documents based on their uses. Then we correlate these types of documents to specific types of KOS according to their degree of structural constraint and activities they aim to serve.

1.0 Introduction

The evolution of work organization models, characterized by an intensification of distant exchanges, the increasing number of coordination and communication tools and of sharing, transmission and back-up systems, results in complex informational environments. In the framework of an ANR project, a new type of Knowledge Organization System (KOS) based on faceted classification is under development, aiming to reduce the cognitive cost of information management tasks in complex digital environments, particularly in working documents management. We are working on a methodology to accompany its deployment and to elaborate relevant facets relating to different trades. In this article, we present a part of this work.

The starting point of this study consists of observations on individual folder organization of documents taken from individual work stations of different research engineers who work in the R&D department of an industrial group. We focus our attention on a particular case of our work-in-progress methodology, concerning the elaboration of facets dealing with document types information, which brings up specific problems. After the development of an empirical typology of observed document types, we propose another theoretical typology to allow the management of document type information. This type of information is essential, yet difficult to process autonomously. Not being of a universal nature, the document type instead aims at representing the different terms of the type according to the context. Hence the document type in a faceted classification is considered a necessary component of document management, whose meaning, through combination with other facets, is rendered unambiguous.

In this article, the theoretical typology we present is established according to document characteristics such as usage, defined as groups they are included in, and for which they represent a support for interactions, and activities for or during which documents
are created. Finally, we propose recommendations to correlate KOS, document uses, and documentarization operations purposes.

2.0 Documentarization and heterogeneous knowledge organization systems (KOS’s)

KOS’s refer to “controlled languages, classification schemes, and to knowledge representation languages from Artificial Intelligence” (Zacklad 2011). In this tool category, Zacklad also includes search engines’ indexes. These KOS’s consist of systems of access to information, knowledge tracking, representation, and filtering systems such as thesauri, classifications, ontologies, and tag clouds. They are most frequently used for documentarization operations on documents which consist of “transcribing or recording a semiotic product on a perennial substrate, which is endowed with specific attributes intended to facilitate the practices associated with its subsequent utilization in the framework of distributed communicational transactions” (Zacklad 2006).

Documentarization is a major issue in knowledge preservation and communication by allowing “(i) to manage them along with other substrates, (ii) to handle them physically, which is a prerequisite to be able to browse semantically among the semiotic content, and lastly, (iii) to guide the recipients” (Zacklad 2006). The stored information related to the documentarization process on a technical level (content), organizational level (coordination), or location aspects (access to documents) accounts for a substantial effort that KOS endorses (Pikas 2007). In addition, we notice that different activities during the trade exercise lead to the production of distinct document types, which are not documentarized with aid of the same KOS. Despite their diversity, the latter differ in structural aspects, and also in content aspects (vocabulary, semantic), though we state that KOS’s present in organizations and their structuring should be correlated to document uses and to the purposes of documentarization operations.

In most organizations, we find frames of reference that define the location where a document should be recorded according to different intentions (record management, sharing, individual use) and to document features (state of document, life cycle, department) to limit informational entropy by controlling document management. The various storage media used according to document features may present heterogeneous KOS and interfaces. Their use appears as an additional cognitive cost regarding those coming from a professional exercise in which the main activity does not consist of information management. (Desfriches Doria and Zacklad 2010).

In fact, KOS diversity, variability of storage media, of activities associated with documents and of document types make document management activities complex. Our findings differ from previous work on typologies by the scope of documents we deal with, in contrast with Zeller (2004), who is interested in all document forms (DTB, Web sites, GIS, multimedia documents, etc.), or to Gagnon Arguin (1998), who focuses her interest on proof documents for record management, or to Alberts (2009), whose work is concentrated on mail and is exploring document gender notion. We limit our studies to digital working documents that we define as individually or collectively produced or handled documents during professional exercise of various trades. The purpose of our approach does not consist of record management, but is more focused on working documents management in a knowledge management perspective.

3.0 Faceted classification

Faceted classification is represented “as a combination of complementary conceptual groups offering the ability to insert varying analysis dimensions on informational objects, to characterize and make access to information easier by offering multiple ways of navigation towards any document” (Mas et al. 2008). The notion of facet often appears as “the most consequent theoretical contribution of the century in information sciences” (Maniez 1999). Faceted classification presents a number of benefits reported in literature. The most common benefits mentioned are expressiveness, flexibility, consistency, and adaptability (Maniez 1999; Ali and Du 2004; Marleau et al. 2008). It has also been recognized by Broughton (2005) to support browsing, navigating, and information researching. This author explains that faceted classification allows browsing (which consists of quickly scanning a corpus to discover its content), thanks to its logical structure and its capacity to express complex or compound subjects. Its structure, which can be combined with user interfaces and multiple access points, enable navigation through a corpus. Finally, information research is supported by progressive filtering based on multiple search criteria (facets) (Broughton 2006), though, according to Kwasnik (1999), one must not overlook the difficulties related to establishing relevant facets, the potential incoherence in inter-facet relations, and in the
visualization of the classification scheme with regard to the internal logic of each individual facet.

3.1 A more flexible approach to faceted classification

Faceted classification is traditionally used, in a formal way, to standardize homogeneous corpus management. Homogeneity is employed here for document types, but also for content aspects. For example, in libraries, document types are almost similar, and contents are described through standards fields as keywords for book subjects. The level of specificity of indexing is established.

By contrast, in working document management, the corpus is heterogeneous in terms of form and amount. We notice that the level of specificity can vary according to specific needs, activities, and amount of produced documents. The content is not necessarily the major indexing requirement; we also meet some specific needs for describing the situation of document creation, like time related information. A study from Pikas (2007) about engineers’ Personal Information Management practices reveals that they do not use the same strategies to retrieve their documents, nor do they remember the same kind of information. This study claims that the most important element while searching for a document is the time dimension, which can be conveyed with differing instances (season, precise date, period, project stage, etc.). The development of relevant facets and of the required level of specificity for the documentarization process is defined in context and in relation to activities, users habits, and volume of produced documents. We don’t recommend any scale, as far as these are principles to be applied in reference to a corpus, a set of activities, a department, or a professional group. Thus the application of the principles of faceted classification in the face of the large diversity of working documents forces us to soften the principle of facets and leads us to reflect on the development of more coherent schemes adapted to diverse situations and actors.

The faceted KOS we develop allows personal faceted classification schemes, without restricting eventual constrained aspects of document description. It emphasizes the flexibility and expressiveness qualities of faceted classification and this way of using it appears as a less strict approach of faceted classification than the traditional ones. From our point of view, users or document creators are considered the most relevant people to index their documents, thus we are developing this methodology for designing faceted classification adapted to all contexts in organizations. In our preliminary study, we notice that documents do not imply identical uses according to different trades, it is not therefore necessary for them to be described in the same terms, in a constrained way by all actors in an organization.

Consequently, our approach to faceted classification which allows it to be fed and developed on the fly, is bottom-up. We can compare it to Vickery’s opposition (1960) to mechanical and constrained implementation of fundamental categories to a subject. These categories should be used as a guide for suggesting potential characteristics that should not be ignored. (La Barre 2010).

4.0 Proposal of empirical and theoretical document typologies

Handling questions about document types leads us to focus our interest on the notion of facet and to confront problems mentioned before by Kwasnik (1999). The choice of relevant facets and the necessity of consistency between facets are influenced by more ancient techniques such as development of lists, taxonomies, or typologies. By typology, we mean analysis and description of typical forms of a complex reality, allowing classification. For our concerns, we need to find division criteria, or dimensions of analysis, from which we can develop a description of empirical complex data, to eventually transfer it to the development of our faceted classification.

4.1 Empirical typology of documents

The theoretical typology of documents presented in part 4.2 represent a means to avoid an increasing number of document types in faceted classification. In fact, during a deep study of folder organization on individual workstation of two research engineers from the R&D department of an industrial group, we noted more than 110 document types which make up our empirical typology. The latter already constitutes a reduction in the actual complexity of observations (Coenen-Huther 2007), given that we found several occurrences of the same document type in folder hierarchies due to the fact that workers are involved in several projects simultaneously with varying roles according to the project.

This empirical typology corresponds to the systematic listing of instances of document types, which we have reduced to a simplified form.
The figure (Table 1) presented above is an extract from the empirical typology. We can hardly accommodate 110 values for a facet with our purpose of reducing cognitive costs of information management tasks, thus we have focused our interest on other dimensions of analysis, such as document usage.

### 4.2 Theoretical typology functions of documents uses

Our theoretical typology is developed from the viewpoint of document usage, which depends, according to us, on groups involved in creation or utilization of these documents and on the purpose of a worker’s activity considered in its entirety and to be seen in the global organization.

In the following table (Table 2), purposes are mentioned in the frame of document creation, as our goal is to enable document management rather than record management.

According to Marradi (1990), this typology, which could also be qualified as an extensional classification, originates with an item set (the document types mentioned in the empirical typology), on which we apply division criteria (purpose of activity and groups types interacting with documents). These criteria are applied to items on the base of property similarity in the item set. Thereby, empirical document types owning the same properties are grouped in a new theoretical and more abstract type.

It can be useful to notice that this typology can potentially be applied to all departments of an organization. For instance, in a Human Resource department, the purpose of the activity labeled as “accomplishment of mission in the frame of projects” can lead to pro-

<table>
<thead>
<tr>
<th>Documentation</th>
<th>Balance Sheet</th>
<th>Case scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Course</td>
<td>Export</td>
<td>Need analysis</td>
</tr>
<tr>
<td>Article</td>
<td>Spécifications</td>
<td>Quality follow-up</td>
</tr>
<tr>
<td>Visual guidelines</td>
<td>Requirements</td>
<td>Review</td>
</tr>
<tr>
<td>Methodology</td>
<td>Trade proposal</td>
<td>settings file</td>
</tr>
<tr>
<td>Thematic bibliography</td>
<td>Test report</td>
<td>Data model</td>
</tr>
<tr>
<td>Transverse bibliography</td>
<td>Appendix</td>
<td>Process description</td>
</tr>
<tr>
<td>Benchmark</td>
<td>Algorithm test</td>
<td>Synthesis note</td>
</tr>
<tr>
<td>State-of-the-Art review</td>
<td>Data retrieval</td>
<td>Working document</td>
</tr>
<tr>
<td>Administrative document</td>
<td>Notes</td>
<td>Preliminary study</td>
</tr>
<tr>
<td>Interview</td>
<td>Thesis</td>
<td>Transverse document</td>
</tr>
<tr>
<td>Study</td>
<td>Letter</td>
<td>Outline note</td>
</tr>
<tr>
<td>Screen shot</td>
<td>Tutorials</td>
<td>Mock-up</td>
</tr>
<tr>
<td>Report</td>
<td>Opportunity study</td>
<td>Consortium agreement</td>
</tr>
<tr>
<td>Presentation</td>
<td>Opportunity Note</td>
<td>Running document</td>
</tr>
<tr>
<td>General bibliography</td>
<td>Need assessment study</td>
<td>Dashboard</td>
</tr>
<tr>
<td>Contribution</td>
<td>Form</td>
<td>Inspect form</td>
</tr>
<tr>
<td>Training</td>
<td>Standard</td>
<td>Contract</td>
</tr>
<tr>
<td>Budget</td>
<td>Input Data</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Extract from the empirical typology

<table>
<thead>
<tr>
<th>Types of Groups</th>
<th>Individual activities</th>
<th>Execution of mission in frame of Projects</th>
<th>Coordination</th>
<th>Contribution to the Trade</th>
<th>Record Management</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project</td>
<td>Document of collaborative work</td>
<td>Project monitoring document</td>
<td>Auxiliary resource document</td>
<td>Document of collaborative work (Final version)</td>
<td></td>
</tr>
<tr>
<td>Trade</td>
<td>Auxiliary resource document</td>
<td>/</td>
<td>Trade document</td>
<td>Trade document (Final version)</td>
<td>Referential document</td>
</tr>
<tr>
<td>Corporate</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Document typology functions of professional activities purposes and types of groups
duction of document types as “contracts.” These documents will be considered, in the context of this activity, as the type “Document of collaborative work,” but could also belong to the category “referential document” from the viewpoint of people from other departments of the organization. The types of groups mentioned in this theoretical typology come from the approach of Zacklad (2007). We assume that this theoretical typology can compose an adding marker for users in the stage of developing document typologies for creation of faceted classification. This can also eventually be a classification principle for consistent faceted organization within a trade.

4.3 Definitions of types from the theoretical typology

In this typology, types are nor definitive nor exclusive. For instance, a document can move from a type “Document of collaborative work” over to an official version for record, and another document of type “Document of collaborative work,” like a data model, can become a “Trade Document” in other situations.

Individual work document: These documents correspond to an individual work activity, aside from any work group, or for documents created in autonomous ways, for preparing to share with a working group. For example: notes, diagram

Document of collaborative work: These documents are written collaboratively, within a group where the work of individuals is highly dependent of other workers’ work, as is frequently the case in project organization. For example: State-of-the-Art-review, requirement specifications

Project monitoring document: This type of document is used to organize activities within projects over time and organizational aspects (coordination). For example: dashboard, schedule

Trade document: These documents are collaboratively written by community of practice members or professionals from the same trade and are used, individually or collectively, for professional activities; they can describe good practices; the type of discourse is often prescriptive. For example: procedure, operating procedure, good practice guide, recommendations

Auxiliary resource document: These documents are often completed, reused in other departments of organizations, or for other projects, or by other professional groups than those who produced them. They are about knowledge capitalization. They are taken up and undergo a revival of interest for activities or projects other than those during which they have been produced. They could be compared to documentation, but we distinguish them because they are internally produced. For example: maintenance documents reused in context of repairing.

Referential document: These documents can be equally accessed by all organization members. They consist of document models, formal descriptions of projects, forms, documents of activities planning of a department. They are not specialized on professional activities. For example: instructions about record management or data backup, consortium agreement, contract, organization chart, visual guidelines

External documents: In every dimension of professional exercise and in almost every case of document production, workers need some documentation. These documents come from information research activities, from external sources.

Record document: Final versions, official versions of individual work documents, documents of collaborative work, trade documents. For example: Deliverable

The following categories of Individual work documents, Documents of Collaborative Work and Trade documents can belong to a broader category from Zacklad (2006), labeled as DofA (Document For Action). These DofA are characterized by their extended state of incompleteness, their perenniality, their fragmentation, their rapid circulation, by the fact that they are produced by different authors and by the non-trivial argumentative relationships between the document fragments. (Zacklad 2006). For Zacklad, DofA corresponds to various devices: textual file or annotated drawings, forum systems, blog systems or wikis, messaging systems, etc. (Zacklad 2007), while we are only focused on working documents in the frame of professional exercise.
The diagram (Figure 1) presented above illustrates that the core documents produced or handled comprises the category of Document of collaborative work. In fact, the Individual Work documents and External documents often contribute to the drafting of Document of Collaborative Work. Trade documents also frequently appear as contributions to this type of document and vice versa. Auxiliary Resource documents generally come from the category of Document of Collaborative Work and also become resources for the drafting of this latter type of document. Lastly, Project Monitoring documents and Referential documents are used to organize the drafting activities of Document of Collaborative Work. Thus, it is not surprising to note that this category gathers the most important empirical types.

4.5. Refinement with activities

One facet containing 40 values is not manageable. It appears necessary to apply a new categorization criteria. We chose the activity element, in which the specificity level can vary in terms of functions of needs, numbers of documents, and degrees of precision needed. Our tool allows the creation of activity contexts for grouping facets with relevance. This enables documentarization with an adaptable level of specificity functions for user needs, in which the functions of the prevalence of certain activities within a trade vary.

If we develop a faceted classification with activity-based contexts, we may find a facet in each context representing specific document types frequently produced during each activity. Thus we can detail document types comprising the Document of Collaborative Work category.

As observed, activities within our KOS have several roles. First, they are a means of grouping facets in a relevant context. Second, they improve information allocation in facets when the number of values is too high by refining the facets’ content, while maintaining consistency in the classification scheme.

The table (Table 3) proposed below is an extraction of reclassifying operations of the Document of collaborative work category functions of specific activities. According to this example, we notice that an acceptable amount of values of facets is created in reference to specific activities. For a facet concerning the preliminary studies documents, the label could be “Preliminary Studies Specific documents.” The executed choice consists of fragmenting document types in reference to activities during which they are produced.
5.0 Recommendations for types of KOS according to document types and management of information activities’ purposes

As mentioned above, we recommend that KOS’s used in organizations and their degree of structural constraint should be correlated to document uses and to documentarization operations’ purposes. Management of information activities and especially for documentarization can be enumerated in a broad outline as follows: applying indexing instructions for record keeping with formal KOS’s, systematic and scalable working documents organization with medium formalized KOS, and tagging of individual work documents through informal KOS. The degree of structural constraint of KOS is related, itself, to document types that are possibly documentarized with this KOS, and storage medias are associated to these features.

We propose to make some recommendations about KOS types functions of theoretical document types and documentarization operations’ purposes. In the table (Table 4) below, KOS’s degrees of structural constraint are correlated to the latter document typology. In addition, we notice that storage media associated with documentarization activities depend on the purposes of these operations and, to an extent, on the public they are addressed to.

6.0 Conclusion

Through a study of document types for developing faceted classification, we recommend degrees of structural constraint for KOS’s used for documentarization of working documents.

Our tool, the flexibility of which has been mentioned before, allows us to apply varying degrees of structural constraint of KOS’s to faceted classifica-

Table 3. Extract from reclassifying one theoretical type of document by specific activities
tion, although it was first designed for a top-down approach for document management, for knowledge management.

The interest in considering activities in the creation of faceted classification lies in the opportunity to make the specificity degree for the classification variable, thus for indexing and then for retrieval. Users’ priorities differ within a department, as does the volume of documents produced during the execution of professional tasks. We assume that the possible variation of degrees in specificity in information management tasks reduces the cognitive costs implied by those activities. Considering activities also allows for fragmenting facet values in several distinct facets, since their amounts might potentially be too large.

Faceted classification makes information management easier by providing multi-point-of-view access to documents. One can remember heterogeneous elements for retrieval, thus, if the searched documents have been indexed by the means of faceted classification, one can recognize potential elements used for the documentarization in facets. Stakes related to graphic interfaces for presenting faceted classification are involved in the efficiency and the success of this kind of system.

Note

1. National Agency for Research

References


Including Authorial Stance in the Indexing of Scientific Documents

Viviane Clavier* and Céline Paganelli**

* Université Stendhal – Grenoble 3, Institut de la Communication et des Médias, 11 av. du 8, mai 1945, BP 337, 38434 Échirolles, <Viviane.Clavier@u-grenoble3.fr>
** Université Paul Valéry, ITIC- Département Information Documentation, Route de Mende, 34 199 Montpellier cedex 5, <Celine.Paganelli@univ-montp3.fr>

Viviane Clavier is Associate Professor in Information Science & Communication Studies at Grenoble3 University. She is a member of Gresec (research group in Information and Communication Studies, Grenoble, Stendhal University). Her research interests revolve mainly around information retrieval and indexing of scientific and professional discourse.

Céline Paganelli is Associate Professor in Information & Communication Studies at Montpellier 3 University. She is a member of Gresec (research group in Information and Communication Studies, Grenoble, Stendhal University). Her research interests revolve mainly around information uses in professional context. She has studied scientists, bankers, and researchers’ information practices, in particular.


ABSTRACT: This article argues that authorial stance should be taken into account in the indexing of scientific documents. Authorial stance has been widely studied in linguistics and is a typical feature of scientific writing that reveals the uniqueness of each author’s perspective, their scientific contribution, and their thinking. We argue that authorial stance guides the reading of scientific documents and that it can be used to characterize the knowledge contained in such documents. Our research has previously shown that people reading dissertations are interested both in a topic and in a document’s authorial stance. Now, we would like to propose a two-tiered indexing system. Dissertations would first be divided into paragraphs; then, each information unit would be defined by topic and by the markers of authorial stance present in the document.

Received 7 March 2012; Accepted 7 March 2012

1.0 Introduction

This article will focus on the indexing of scientific documents in connection with the context of their use. Working in the field of specialized information, we believe that knowledge is the product of an encounter between the data contained in a document, a user in search of information, and a specific context. Context may refer to an organizational, technical, or human environment. Knowledge is, as such, understood as both the result and interpretation of data in connection with users, their activity, and a given context. As such, issues surrounding data collection and the representation of knowledge must take into account the nature of a document, the user, and the user’s main activity, which we will refer to here as the “context of use.”

The research presented in this article follows from exploratory research conducted on doctoral theses readers, the outcome of which was published in the Les Enjeux de l’Information et de la Communication electronic journal (Clavier and Paganelli 2010). Our goal was to assess the relevance of the notion of stance in reading and annotating for indexing and
knowledge representation purposes. As an extension of our previous research, we would now like to more specifically point out how linguistic and cognitive knowledge, in connection with stance, can be used to improve access to information.

Stance is not a linguistic category per se, but the term is used to designate a series of linguistic processes typical of scientific writing. The large amount of research conducted in the context of research projects is proof of a strong interest in scientific discourse. For example, the Norwegian KIAP project (Kulturrell Identitet i Akademisk Prosa) and the French Scientext project focused on scientific writing. In the latter project, specifically, stance refers to the linguistic processes that reveal “an author’s singularity, their specific contribution – the justification behind their scientific approach – and the author’s reasoning, that upon which the research is based, the proof used, the logical relationships it establishes – the quality of the scientific analysis.” We believe that an author’s stance is a driving notion that guides the consultation of scientific documents and is also central to describing their content; as such, we feel that stance needs be a full-fledged part of the indexing process for doctoral theses.

We shall begin with a presentation of the theoretical footing on which our approach is based. Then we will show how the notion of stance is mobilized by users when consulting scientific documents. Finally, we will formulate a certain number of proposals for indexing and the representation of knowledge conveyed in scientific discourse.

2.0 Theoretical Framework

Our approach is part of a body of research from the information and communication sciences. Given this disciplinary rooting, we have not addressed the representation of knowledge in terms of the formalization of data in the information technology sense of the term; it is not understood as the development of organizational systems in the knowledge management sense either, but it does rely on the description of methods which allow us to draw out data that fuel systems of knowledge representation. There are two trends in the information sciences which differ in how they understand information: recorded knowledge and communicated knowledge. Hubert Fondin has argued that information is part of a process of exchange and sharing, of finalized communication, in a specific context or social system (Fondin 2001, Fondin 2005), and information is, as such, understood as communicated knowledge. Conversely, Yves-François Le Coadic has posited that “information is knowledge recorded in written, oral or audio-visual form on a spatial-temporal medium” (Le Coadic 1994, 6, translated here). For Le Coadic, information is thus understood as recorded knowledge. Our research tends to identify with the first approach since we believe that knowledge exists when there is interpretation, assimilation by an individual and when it is connected to a universe of defined knowledge. Further, we believe that knowledge is constructed by individuals according to the context of use.

This so-called context of use is thus fundamental, both theoretically and methodologically speaking. A lot of research over the past ten years has shown that context has a strong influence on information activity. Brigitte Guyot (2002) has notably shown how information activity is becoming increasingly important in professional contexts. Factors from all levels are involved and influence informational activity—affective states (Kuhlthau 2004) or the specific constraints of a task (Järvelin and Ingwersen 2004)—and a lot of research has focused on information habits in specific professional contexts (Cheuk 1999, Miranda and Tarpapoff 2007, Staii et al. 2006, to name but a few), thus considering that an information activity is affected by context and the activity underway (Bartlett and Toms 2005, Li and Belkin 2008).

This approach has consequences for the methodology behind data collection. We believe that, in some respects, context needs to be taken into account when defining how documents should be processed. This perspective places us within the actor-oriented paradigm (Polity 2000, Chaudiron and Ijadadiene 2002) which includes research that sees information as an interpretive process and that underscores the importance of the concept of context in informational activities (see notably Fidel and Pejtersen 2004, Byström 2007).

We believe that context of use is defined by three variables that have been widely addressed by research, either independently or in a combined manner, and under many different albeit sometimes similar designations, such as the notion of “task,” for example, commonly found in English language research in Library and Information Science (Järvelin and Ingwersen 2004, Byström 2007, Huvila 2008):

- Cognitive factors related to individuals in the context of their work (individual factors: expertise, know-how, the universe of knowledge, etc.);
- Factors related to a person’s professional activity (main activity for which a user is conducting an in-
formational activity, consults or is looking for information in documents; an activity that occurs within a socio-organizational context);

- Factors related to an application (systems, sources of information, documentary genres, specialty fields, etc.).

It is a combination of these three factors that allows us to gather information to represent knowledge. We chose to focus on three sources for the collection, identification, and interpretation of knowledge in order to represent it: documents, users, and the motivations that push an individual to look for information and consult documents. This methodological stage required us to collect “traces,” a term we used to designate data collection methods that allow a corpus to be compiled. Our corpus was defined according to the three sources mentioned above and drew on:

- Documents consulted by users in a professional context, if possible at their place of work. As Dominique Cotte has noted, a document is a very specific object since it is not “data” but rather a “constructed product” resulting from the combination of “signs, alphabetics, images, diagrams, [that] can form texts, supported by documents, which may or may not contain information” (Cotte 2004, 31-32, translated here).

- Traces of use or more broadly the “traces of activity” found on such documents (Flon et al. 2009), such as annotations left by a reader on a consulted document or all of the “sources of marking” automatically collected and “redocumented” (Yahiaoui et al. 2011) to explain the “human and social context of activities.” There are various methods for collecting such traces: automatic collection recorded following a computerized action; semi-structured interviews that aim to clarify motivations, the reasons behind the choice of one document, or part of a document over another; and collecting verbal protocols that aim to make subjects “speak out loud” when consulting a document, for example.

This approach was implemented in different contexts, all of which involved a professional situation with users who needed to accomplish a main activity (computer maintenance, writing a thesis, etc.) for which they conducted an information activity. Our previous research conducted in professional contexts (Paganelli and Mounier 2002, Clavier and Paganelli 2010) has shown that information activity is secondary and subordinate to one or more main tasks (preparing a course, doing computer maintenance, etc.). This leads to different types of reading which are driven by the reader’s goals. Regardless of these goals, reading in a professional context is generally fragmented and non-sequential and involves a large amount of physical and cognitive activity (copy-pasting, underlining, annotations) that leave numerous traces of an individual’s informational activity (Hochon and Jacobini 1994, Mille 2005). In work contexts and depending on the sector, the documents we examined were maintenance manuals, legal texts, medical reports, theses and research articles. Different studies have shown that such documents contain formal characteristics (linguistic and structural) that can be used to improve automatic processing in order to represent the knowledge contained in a document (Péry-Woodley and Scott 2006, Poudat et al. 2006, Couto and Minel 2007).

3.0 Stance as a common thread in the consultation of theses

The way theses are consulted changed a lot when they became available online. The consultation of such documents remains marginal on paper, but has greatly increased for digital versions. Since 2000, a number of projects and efforts to disseminate electronic versions of theses have emerged, and such initiatives beg us to think about access methods and the principles of indexing. The question is not new in and of itself. Sylvie Lainé-Cruzel has defined an information system piloted by user profiles for consulting scientific documents (Lainé-Cruzel 1999), and other research has focused on access to French theses in digital libraries (Abascal-Mena and Rumpler 2007). In the first case, however, access to sources is filtered by the profiles, which is fairly restrictive; and, in the second case, the focus is placed on the semantic content of the document via the extraction of concepts, which limits access to the document’s terminological dimension.

The experiment we conducted has been described in Clavier and Paganelli (2010); it was conducted in three parts. The first phase involved observing the thesis reading habits of ten doctoral candidates in information and communication sciences. Then we questioned them about the criteria they used when selecting theses, and we gathered their comments about the passages of text considered important. We then created a corpus of textual fragments (the passages read) to which we added written annotations from the different media (the actual theses, files, post-it notes, etc.). We also collected oral comments from readers regarding either their consultation strategy or the pas-
sages of text selected. These data were then entirely transcribed and comprised the corpus to analyze.

Among the observed results, it appeared that the consultation of theses by doctoral candidates occurs in a professional setting, in the context of their own research. This type of use corroborates what has been observed in other professional environments (Paganelli and Mounier 2009; Statt et al. 2006): a non-contiguous, often partial reading that leads to an infinite number of experiences influenced by the specific tasks at hand (seeking a definition, problematization, etc.). We observed that approaches to reading differed depending on the number of years a candidate had been preparing their thesis: while readers first seek to “learn the landscape” (become familiar with authors, schools of thought, grasp the terminology, etc.), they later aspire to situate themselves (quoting one author rather than another, identification with a school of thought, adopting their own terminology). As such, while topics are useful for choosing a document or the parts of a thesis to be consulted, it is the metadiscursive elements that reveal the author’s stance which truly guide reading.

Our analysis of the corpus allowed us to identify the indicators of stance and interpret them. In doing so, the annotations added by readers and the oral comments associated with each passage of text allowed us to see how readers understood the documents they consulted. Such personal traces are a means for the reader to take possession of a document and interpret its content (Mille 2005). We analyzed 158 text fragments: of these, 129 had visual markers (underlining, highlighting, etc.); 47 contained annotations (notes, abbreviations, keywords, symbols); and 148 were commented on orally. The annotations and comments allowed us to identify two types of indicators in the fragments. The first occurred at the discourse level; the second at the textual level.

In the first case, the indicators collected were evaluative, axiologic, and from epistemic and evidential categories. We, as such, found the linguistic markers mentioned in Grossman and Wirth (2010), Boch et al. (2007), and Rinck (2010), although there were fewer categories than in their research. In the second case, the indicators collected allowed us to localize statements according to their position in the document. We thus agree with Alain Berrendonner (1997) who has argued that “meta-discursive pointers” exist which are deictic (“here, see over”), text extracts (“in the first section”) or even imprecise locations (“in this passage”) and for whom a document is a “vectorized textual space.”

4.0 Suggestions for indexing theses by stance

4.1 Points of view, facets and terminological variations in stance?

Unlike the notion of “point of view,” which finds resonance in information and documentation and amongst researchers in linguistics and computer science working on textual data (corpora, databases, the internet), the term stance is not commonly used in information science. In the context of information and documentation, indexing using Shiyali Ranganathan’s faceted classification system dates back to the 1950s. Facet analysis is not, strictly speaking, an enunciative approach that follows the author’s point of view, but rather it allows different points of view to be expressed about an object (Salvan 1962). Without reference to the famous classification system, Bachelin Ralason (2010) has also employed the term facets when seeking to provide a multi-faceted representation of a document using several ontologies (ontology of topic, field, task, etc.). In this case, these representations involve the thematic content of a document, as well as its application context. Research conducted in the context of the RAP2 project has also underscored the interest of searching for information by point of view, thus allowing the user to focus on specific approaches to a concept. A whole collection of terms, called linguistic markers (Laublet et al. 2002), is associated with each point of view. To conclude this quick overview, let us mention research based on corpus linguistics which addresses scientific writing more specifically. The concept of point of view is central in pointing up an author’s scientific rhetoric (Teufel et al. 1999) and their enunciative position (Tutin et al. 2009) based on language. Such language markers are discontinuous, rooted in discourse or meta-discourse, and, as Ho-Dac and Péry-Woodley (2008, 3) have argued, they should not be confused with segmentation markers, but rather are indicators that “help nourish a relationship of continuity or discontinuity between two segments.”

4.2. The triangular approach to stance

Our previous research into the indicators of stance pointed up two important limitations: first, there is a
great diversity of markers that refer to numerous semantic categories which occasionally intersect and are difficult to grasp. Secondly, the dissemination of indicators throughout a text makes all attempts at indexing via this approach impossible. As such, we established that it is best to limit the notion of stance to three categories of markers that must simultaneously be found in a sentence or, at most, a paragraph. These categories set the triangular boundaries that delimit a stance’s field of application: 1) Expressions that reveal a judgment or an author’s subjective comments (agreement, mitigation, criticism, consensus, etc.); 2) expressions that name a topic (terms, concepts, propositional content, etc.); and 3) Expressions that mention the given environment (or give a reference mark)—this can be in discourse (dates, places, references to others, etc.) or in a document (chapter, section, etc.). Here are a few examples that contain indicators of stance. These extracts are part of a thesis read by one of the people interviewed for our research.

– E1. Il demeure cependant indéniable que l’hypertexte est un terme qui fait aujourd’hui partie de notre culture commune. (Ertzscheid, sujet 3)
– E2. Sans point commun apparent avec l’idée de Nelson, il est intéressant de remarquer comment, au point actuel de l’évolution technologique, les deux définitions entrent sans peine en résonance, laissant entrevoir un champ épistémologique à la fois ouvert et complexe dans lequel les associations de l’un font écho aux « dérives » de l’autre. (Ertzscheid, sujet 3)
– E3. Nous défendons dans ce travail la thèse selon laquelle l’hypertexte n’est pas un épiphanomène de nature informatique assimilable ou réductible à l’un des sphères de la réalité qui l’emploie. (Ertzscheid, sujet 3)

Each of these extracts contains the three categories of markers: 1) Expressions that identify a stance (assertion, statement, thesis): “il demeure indéniable que (E1), il est intéressant de remarquer que (E2), nous défendons la thèse selon laquelle” (E3); 2) Expressions that describe a topic: “hypertexte” (E1, E2, E3); and 3) Expressions that allow us to locate a point of view (time, place or angle of approach): “aujourd’hui” (E1), “sans point commun avec l’idée de Nelson” (E2) “En prenant l’angle critique qu’offre l’analyse des hypertextes littéraires” (E3) (document, chapter, imprecise location); “dans ce travail” (E3).

Although we do not yet have any precise data on the efficiency of this model, we have chosen (at first) to recommend the most restrictive model since it requires that three levels of information be simultaneously present. Stance thus has composite status: it combines elements of language that are both indicial and relational and points to markers at different levels (lexical, syntactical). This proposal is based on the localist framework (see Clavier and Paganelli 2010).

4.3. Connection between indexing and practice

The systems that provide access to digitized theses offer various means to search for information: generally, access by structured field (author, title, etc.) and access by content (title, abstract or keyword). Occasionally, it is possible to search the entire text. In order to improve access to information in theses, we recommend including knowledge about authorial stance and connecting it to indexed topics. This representation would involve a twofold indexing process. After segmenting the text, each fragment from the cut-up would be described by both the topics it contains and a label indicating whether or not indicators of stance are present. Such dual indexing would exist on pre-identified and segmented units of information; we believe that paragraphs are the most appropriate basic units for the segmentation and indexing of large documents (Mounier and Paganelli 2003).

On the first level, topics would be indexed according to the structure of the document. This approach has notably been described by Abascal-Mena and Rumpler (2007) with regard to theses; an overview of existing methods for the thematic indexing of long documents like monographs has been done by Lyne Da Sylva (2004).

On the second level, units of information would be characterized according to whether or not indicators of stance are present. When indicators are present, the nature of the stance (critical, agreement, etc.) would be mentioned. The way indexes are structured offers for two possible solutions.

In the first case, indexes by topic and marker of stance would be dissociated; in the second case, one index would contain both sets of information: the topics and whether they do or do not contain stance markers. The first solution would be linguistically more coherent since there would be an index for each level of information. Conversely, the second solution would offer the advantage of listing topics that are or are not modalized. Both types of indexing would allow for research that combines searches by topic and stance; the indexes would need to be designed to be included in the primary document rather than be
separate; they would also need to be designed as reading tools that allow us to manipulate no longer entire documents but rather segments of text. In this respect, our proposals are similar to the recommendations made by Muriel Amar (2004) regarding the nature of indexes needed for new digital media.

5.0 Conclusion

We took as the basis for our research that different approaches to reading scientific documents could be interpreted through an analysis of the traces of informational activity. This methodology allowed us to empirically confirm the relevance of the notion of stance when consulting theses on the one hand and, on the other hand, the interest in associating an author’s “global” point of view (criticism, agreement, consensus, etc.) towards the topics, notions, and concepts addressed in a document. We also suggested representing the infinite number of reading experiences in the form of stable knowledge likely to be represented in indexes. This research needs to be pursued with the systematic collection of markers in order to assess the degree of automation in indexing. This understanding of knowledge is related to indexing as an interpretive process that cannot be imposed by a controlled vocabulary or solely by the text but which is also mediated by the traces of an individual’s use in the context of their work.

Notes

2. Scientext: un corpus et des outils pour étudier le positionnement et le raisonnement de l’auteur dans les écrits scientifiques [a corpus and tools to study authorial stance and reasoning in scientific texts], directed by Francis Grossmann and Agnès Tutin, ANR 2007-2010 http://scientext.msh-alpes.fr
4. In France, the information and communication sciences form a single discipline, which makes them somewhat of an exception.
5. From an internal document produced by the Grenoble sicd2: “the consultation figures for digital theses are impressive. For the 4000 theses available on the TEL/CCSD server, there are over 100 downloads per day, whereas a paper thesis is consulted on average once every ten years” (translated here).
6. Notably at the Lyon 2, Paris 12, Aix-Marseille 2 and Bordeaux 1 Universities and at Insa Lyon.
7. For Alain Berrendonner, a text is an “organized collection of successive utterances accomplished over the course of a discourse” and “vectorized space” is a “schematization of the text as space” [ibid. 221, translated here].
8. This last statement needs to be verified through more in-depth research.
9. TEL (tel.archives-ouvertes.fr/) or Fourier University’s collection of online theses (tel-ujf-grenoble.fr/).
10. Lyon 2’s “Cybertheses” (theses.univ-lyon2.fr) or theses from the Strasbourg universities (http://scd-theses.u-strasbg.fr/).

References


As an enthusiastic promoter and practitioner of interdisciplinary research, Rick Szostak believes that disciplinary boundaries artificially and unnecessarily create obstacles to interdisciplinary scholarship. Classification by phenomena rather than discipline would expose researchers to different methodologies and reduce scholarly duplication, thus advancing research more quickly (2003; 2004; 2008; 2010). Showing the causal links between theories would reveal formerly hidden connections, benefitting both students and scholars. In two volumes, one aimed toward the philosophy of science community (2003), and the other for information science (2004), Szostak partially lays out his own classification scheme that classifies by phenomena, data, method, theory, and practice, along with practical instructions for applying it, including a notational system. He intends the 2003 volume not to be a bibliographic classification, but rather a “map” of the types of human sciences, whereas the 2004 book is meant to be a reference work for use by scholars, in particular graduate students and interdisciplinaryians, to assist them in seeing “full range of possibilities” of phenomena and “what uses particular theories and methods best serve” (2004, xiii). As a point of reference, Szostak (2004, ix) notes that, in his view, “science’ encompasses the natural and social sciences, plus the humanities.” In the 2003 book, he provides a justification of his project and an outline of the categories of phenomena under the concept of “culture” and how they can be organized and linked. In his 2004 work, he specifically describes his 5W approach (Who, What, Where, When, Why), questions used to guide the classification of scientific documents by phenomena, data, methodology and practice.

Szostak picks up on a tradition that originated in the mid-1960’s when the Classification Research Group (CRG) investigated the potential to create a universal faceted scheme, not arranged by discipline. The group believed that disciplines, as used in dominant schemes such as LCC and DDC, resulted in classificatory rigidity hostile to new concepts (Spiteri 1995). The result, though not ever fully realized, was the theory of integrative levels, introduced by Douglas Foskett (a copy of the draft schema can be viewed at http://www.iskoi.org/ilc/crgc.php). The system was partially operationalized by Derek Austin, forming the inspiration for his Preserved Context Index System (PRECIS) (Gnoli and Poli 2004, 154). Other KO researchers have advocated for or attempted to devise ontological classifications based on a similar structure, such as Dahlberg (2008, 163), who applies the integrated levels based on Aristotle’s four levels to her Information Coding Classification (ICC). More recently, the efforts of the CRG has been continued by the advocates of the Integrated Levels Classification (ILC) structure, whose mission, as expressed in the León Manifesto are to be innovative, phenomena-centered, faceted, and reflect the “multidimensional nature of complex thought” (Gnoli and Szostak 2007). Although he is now a frequent collaborator with members of the CRG, particularly Gnoli, Szostak was not associated with the group until after his scheme was created, and the work of none of the CRG members was referenced in either volume.

The central idea behind integrated levels is that reality is divided into fundamental divisions to which all phenomena belong or have qualities; Hartmann’s fu-
damental divisions, for example, are material, organic, mental and social (cited in Gnoli 2007, 169) and Dahlberg’s (2008, 163) nine levels include “Matter and Energy,” “Societal Beings” and “Intellectual Products.” The “integrative” level means that basic levels combine to make a new item, such as steel (part relationship), or, an “aggregative” level creates a new thing where the individual elements retain their original identities (whole relationship) (Gnoli & Poli, 2004, p. 155); for example, a gaggle (whole) is made up of individual geese (individual wholes). These levels can repeat unlimitedly to reach the appropriate level of specificity, but then break down into facets to express the purpose or function of the concept or phenomena (Gnoli 2006, 138). For example, a chair could be considered in terms of several uses or purposes. In a discipline-based classification, those different functions of “chair” would be scattered in several LCC classes, for example, manufacture (TS), depictions in art (ND), and decorative arts (NK). Using integrative levels, purpose and function are included as facets of the search process to ensure that the appropriate use of the topic sends the searcher on the right path (Gnoli & Poli, 2004, p. 158). So, for the example of chair, the basic level might be something like “Matter,” integrating until it reaches “chair,” at which point the subdivisions (facets) would indicate the specific function or purpose of “chair” the searcher requires.

Szostak’s schema has structural similarities to the ILC, but differs in significant ways. Szostak’s (2003) categories of phenomena include culture, non-human environment, genetic predisposition, individual differences, economy, politics, social structure, technology and science, health, population and art (329-335). He then breaks down into second level phenomena and third level phenomena, in a hierarchical structure. He too, envisions a synthetic structure where hierarchically organized phenomena and methods are enumerated, but with the theory and scientific critique facteted rather than function or context (2004, 220). Rather than asking for named theories and methods, Szostak instead uses the 5W questions to analyze the components of each. To break down theories, for example, he sets up a chain where the “who” of theory identifies the agent effecting change, the “what” is what the change agents do, the “why” is why the change agent acted in such a way. He asks whether the behavior is intentional or non-intentional, and whether those agents are individuals, groups, or whether the theory focuses on the interaction between people (2004, 56). He then delineates positivist, interpretivist, and passive (constraint and incentive based) and lays out five types of ethical analyses (2003, 66-67). For notation, he suggests that phenomena be represented by mnemonic letters and facets by numbers from a list. This has been practically developed as a synthetic notation for theory types using ILC phenomena (Szostak and Gnoli 2008).

Over the last few years, a polite debate has been simmering on the pages of the Journal of Documentation and the Journal of the American Society of Information Science & Technology between Szostak and Birger Hjørland over their fundamental approaches toward classification of the human sciences (Szostak 2008, 2010; Hjørland 2008, 2010, in response to Hjørland and Pederson 2005, Hjørland 2009 and Szostak 2008). Instead of the universalist stance taken by Szostak, Hjørland argues for a domain-centric approach based on the epistemological positions of pragmatism, rationalism, empiricism, and ideally, historicism (cf., Hjørland and Hartel 2003; Hjørland and Pederson 2005; Hjørland 2008). In particular, pragmatism dictates that the concepts are defined in accordance with how the information will be used by the audience of the classification. He believes that concepts are the product of the disciplinary domain, and that theories produce definitions of concepts that may differ from each other, even within a domain. Furthermore, particular domains may require multiple classifications to accommodate different user groups or different goals (Hjørland 2008). Hjørland has many followers, mainly because his theory manages the pseudo-poststructuralist problem that has plagued classification, that is, the problem of multiple realities creating differing conceptual definitions based on perspective. Put simply, he rejects the notion of a universal classification and believes that Szostak’s rationalist approach “assumes a kind of universality and stability that is problematic” (Hjørland 2010, 1079). Despite managing multiple perspectives, Hjørland’s theory does not address classification in broader environments that might call for inclusion of multiple domains.

Hjørland argues that Szostak conflates positivist and pragmatic approaches, or, that Szostak does not find problematic the act of “choosing” a viewpoint “at the expense of other views” (Hjørland 2008, 337). Indeed, Szostak claims that his scheme can be “applicable to all people and societies” (Szostak 2003, 44), and that the existence of basic phenomena is not in dispute “across scholarly communities,” but rather, their “relative importance” is what is in dispute (2010, 1076). Szostak believes that scholarly consensus on the definition of concepts to be possible, and where
no consensus exists, at least a middle ground can be found. He claims that “to deny the very possibility of an objective reality” is to “place arbitrary limits on scientific practice” (2003, 77). He follows the Platonic notion that as humans, our imperfect senses prevent us from reaching the truth, but through research, we can get closer and closer to it. Hjørland (2008, 335), on the other hand, believes that “neutralit[y] and ‘objectivity’ are not attainable” and that “Any given classification will always be a reflection of a certain view or approach to the objects being classified” whether it is easily detectable or not. Although in reality, Szostak’s schema attempts universality, he claims that it “can be seen as a postmodern attempt to show how science can deal with complexity,” which he takes to mean “embrac[ing] diversity” (2003, 42). He dismisses postmodern thinkers who believe universal scientific understanding to be impossible, instead taking an optimistic approach that it is indeed possible to manage multiple meanings as intersections of a variety of causal links. He believes that postmodernists can either “give up hope of advancing our understanding so that we can aid society ... Or, they can strive to battle complexity and subjectivity, holding out hope that we can slowly advance our understanding” (2004, 43) with the assumption that a singular truth exists to be discovered.

Szostak’s work goes beyond Cutter’s advisory function in that he believes that classification should be overtly evaluative; that is, he wants his classification to “reflect some theoretical order” that would help “identify strengths and weaknesses of different types of science” (2004, 2). Theories should be “evaluated on their merits according to...criteria. Theories that are composed of illogical or unrealistic components should be highly suspect” (2003, 80). He continually invokes Aristotle’s Golden Mean to justify his project and decisions, by which he means that “a belief that the truth generally lies somewhere between extreme positions” (2004, 247). However, with the Golden Mean, Aristotle meant the “extremes” to be vices, and that the mean is not a universal truth, but rather an ethical emotional reaction that shifts depending on the context, so much so that no universal rules can be made to guide its use (Nicomachean Ethics). In terms of classification, including the “extreme” ends is not a defense of relativism as much as an acknowledgment that warrant has some role in classification. It appears that extreme views are normed out of Szostak’s schema, which erases important scholarship to arrive at the middle ground that he deems “correct” (2004, 16). However, the goal of document classification is access, so as disagreeable as some extreme views may seem, they may be still need to be accessible.

The Golden Mean approach in this case simply obscures alternate viewpoints with the chosen view based on the judgment of one, based on surface knowledge of the phenomena. The explanations he provides buttress Hjørland’s (2008, 335) point that “Any work on any subject is always made from a point of view.” For example, in his discussion of sexual preferences, he writes, “Not surprisingly, suggesting genetic determinants of gender-specific cultural behavior can be controversial. This is regrettable” (154-158) and goes on to cite six male zoologists, psychologists and anthropologists, no women, and no scholars from gender studies to support his position. In terms of classification, the subordination or marginalization of minority views, people and ideas in purportedly universal classification scheme has been the target of criticism from A.C. Foskett (1971) to Olson (e.g., 1998, 2002), along with other KO scholars. These scholars argue that the ostensibly objective truth has been constructed by the powerful with the intention of constructing “reality” as a means to maintain power for the dominant group. Szostak acknowledges that some might consider existing scholarship “suspect due to the past domination of scholarship by white middle-class males” (2004, 44) but as he describes his decision-making process in detail, he shows his classification is a series of decisions. It is the opinion of this author that judgments do not make universal truths.

Furthermore, the logistical issues associated with such a classification cannot be disregarded. Szostak (2004, 34, emphasis his) writes that his schema shows that “human science must engage thousands of links among thousands of phenomena,“ which means that the catalogue must be able to break down the theories and phenomena using Szostak’s method, with a critically thorough understanding of the work, of theories and methodologies, and of the links to other phenomena. Additionally, Szostak (2004, 234) acknowledges that the classification as it exists will cause huge recall in retrieval because so many documents will be associated with each component part, which then “can only be solved by scholarly efforts to summarize research” through meta-analyses and survey articles. It is unclear, however, how “Writers of surveys [can] overcome the tendency to stress their favored view” or even how coverage summarizing all human science can be achieved or maintained at all (234).

Even with the above points of criticism, Szostak has thought creatively, with the hope of improving the scholarly process rather than merely critiquing the ex-
existing classification schemes. His ideas have benefited from collaboration with the CRG, and some components of the scheme could supplement existing classification practices. His goal of enhanced interdisciplinary research certainly is valuable and ambitious, and the dialogue his work has opened up can help advance interdisciplinary classification.

References


Reviewer: Melodie J. Fox

Affiliation: University of Wisconsin-Milwaukee, School of Information Studies

E-mail: msjoblad@uwm.edu
KO Literature now searchable online

Since its foundation, ISKO has been producing bibliographical bulletins of the literature published in the field of knowledge organization (KO). These followed Dahlberg’s major International Classification and Indexing Bibliography, in two volumes covering years 1950-1982, with subsequent additions. The KO Literature bulletins, compiled by the knowledgeable and patient work of their editors Ingetraut Dahlberg, then (1997-2006) Gerhard Riesthuis, then (2007-present) Ia McIlwaine, were published regularly within this journal.

Since 2009, ISKO decided that the bulletins would be moved to a digital format and be available freely through our website <www.isko.org/lit.html>. This allowed prof. McIlwaine to start feeding a structured bibliographical database, set up in 2010 by me as the new webmaster, instead of formatted text files.

The structure of bibliographical records in the database reproduced that of records in the past bulletins, in view of interoperability with the back data. Thus, each record includes an identifier (originally counted from 1 onwards each year, then from 1 without any reset since the bulletins have moved online) and a classmark based on Dahlberg’s classification of KO Literature (Knowledge Organization, 20 (1993), n. 4, p. 211-222; republished at <www.isko.org/scheme.php>, 2011) which applies the principles of her Systematifier.

Although this classification scheme was conceived several decades ago, thus needs some adaptations in order to class contemporary topics, its continued use of for quite a long time is a huge advantage as it offers the potential for unified subject search. Enabling easy access to consistent bibliographical information in our field can promote the knowledge of its valuable heritage on a wider basis, especially among different communities of researchers dealing with related although traditionally separated topics, each getting more and more relevant for the others in our telematic era.

With this idea in mind, I have started to look for possibilities of converting back data so to upload them to the database. An agreement between Ergon-Verlag and ISKO presidency allows us to make them freely available in our website. Digital versions of the past bulletins from 1997 to 2008 have been kindly provided by their editors. After some consideration of possible solutions, a skilled collaborator has been identified in Pietro Fezzardi, student in Mathematics at the University of Pavia, who has been charged by ISKO of the data analysis and conversion work. This stage has been completed successfully last April, and the data have been uploaded in the database and merged with the current ones.

Since May 2012, the KO Literature section of our website allows for multi-field search in the integrated database from 1997 to present. Records can be extracted by class, author, word in title, year, and language. These fields can be combined creatively in order to get from a broad sample such interesting information as: how many papers on UDC in French vs. in Spanish are recorded, by comparison to the same ratio for papers on thesauri?

As some members had rightly remarked, the classmark for each record was not too useful as long as it was displayed without its verbal meaning -- a fact that we as knowledge organizers are even more aware of than our users. While in printed bulletins it would have been complex to show this information for layout and space reasons, in the Web environment this is now possible. So the classmarks are now automatically parsed by a PHP script, fitting the syntactical features of the Systematifier-based scheme for KO, which allows for synthesizing notation in several ways. The verbal headings of the corresponding isolates are thus identified in an almost perfect way, and displayed in dark orange colour on the top-right corner of each record, from where they can be selected for further navigation by subject.

No doubt, details of both data themselves and the search and display interface can be further improved. KO researchers are encouraged to spread the news of their availability, to try it and to report problems and suggestions to webmaster@isko.org.

Let’s hope that this service, building on the work of past and present compilers, will be only the start of a successful international database of KO literature.

Claudio Gnoli
gnoli@aib.it
KNOWLEDGE ORGANIZATION
KO
Official Bi-Monthly Journal of the International Society for Knowledge Organization
ISSN 0943 – 7444
International Journal devoted to Concept Theory, Classification, Indexing and Knowledge Representation

Publisher
ERGON-Verlag GmbH, Keesburgstr. 11, D-97074 Würzburg
Phone: +49 (0)931 280284; FAX +49 (0)931 282872
E-mail: service@ergon-verlag.de; http://www.ergon-verlag.de

Editor-in-chief (Editorial office)
Dr. Richard P. SMIRAGLIA (Editor-in-Chief), School of Information Studies, University of Wisconsin, Milwaukee, Northwest Quad Building B, 2025 E. Newport St., Milwaukee, WI 53211 USA.
E-mail: smiraglia@uwm.edu

Instructions for Authors
Manuscripts should be submitted electronically (in Word or RTF format) in English only via email to the editor-in-chief and should be accompanied by an indicative abstract of 100 to 200 words.

A separate title page should include the article title and the author’s name, postal address, and E-mail address, if available. Only the title of the article should appear on the first page of the text.

To protect anonymity, the author’s name should not appear on the manuscript, and all references in the body of the text and in footnotes that might identify the author to the reviewer should be removed and cited on a separate page. Articles that do not conform to these specifications will be returned to authors.

Criteria for acceptance will be appropriateness to the field of the journal (see Scope and Aims), taking into account the merit of the contents and presentation. The manuscript should be concise and should conform as much as possible to professional standards of English usage and grammar. Manuscripts are received with the understanding that they have not been previously published, are not being submitted for publication elsewhere, and that if the work received official sponsorship, it has been duly released for publication. Submissions are refereed, and authors will usually be notified within 6 to 10 weeks.

The text should be structured by numbered subheadings. It should contain an introduction, giving an overview and stating the purpose, a main body, describing in sufficient detail the materials or methods used and the results or systems developed, and a conclusion or summary.

Footnotes are not permitted; all narration should be included in the text of the article.

Italics should not be used for emphasis. Em-dashes should be used as substitutes for commas. Paragraphs should include a topic sentence and some developed narrative. A typical paragraph has several sentences.

Reference citations within the text should have the following form: (author year). For example, (Jones 1990).

A citation with two authors would read (Jones and Smith, 1990); three or more authors would be: (Jones et al., 1990). When the author is mentioned in the text, only the date and optional page number should appear in parenthesis – e.g. According to Jones (1990), ...

References should be listed alphabetically by author at the end of the article. Author names should be given as found in the sources (not abbreviated). Journal titles should not be abbreviated.

Multiple citations to works by the same author should be listed chronologically and should each include the author’s name. Articles appearing in the same year should have the following format: “Jones 2005a, Jones 2005b, etc.” Issue numbers are given only when a journal volume is not through-paginated.

Examples:


Illustrations should be kept to a necessary minimum and should be embedded within the document. Photographs (including color and half-tone) should be scanned with a minimum resolution of 600 dpi and saved as jpg files. Tables and figures should be labeled with a detail and title at the bottom, and all columns and rows should have headings. All illustrations should be cited in the text as Figure 1, Figure 2, etc. or Table 1, Table 2, etc.

Upon acceptance of a manuscript for publication, authors must provide a wallet-size photo and a one-paragraph biographical sketch (fewer than 100 words). The photograph should be scanned with a minimum resolution of 600 dpi and saved as a jpg file.

Advertising
Responsible for advertising: ERGON-Verlag GmbH, Keesburgstr. 11, 97074 Würzburg (Germany).
© 2012 by ERGON-Verlag GmbH.
All Rights reserved.

KO is published bi-monthly by ERGON-Verlag GmbH.
– The price for the print version is € 198,00/ann. including airmail delivery.
– The price for the print version plus access to the online version (PDF) is € 229,00/ann. including airmail delivery.

Knowl. Org. 39(2012)No.4
Scope

The more scientific data is generated in the impetuous present times, the more ordering energy needs to be expended to control these data in a retrievable fashion. With the abundance of knowledge now available the questions of new solutions to the ordering problem and thus of improved classification systems, methods and procedures have acquired unforeseen significance. For many years now they have been the focus of interest of information scientists the world over.

Until recently, the special literature relevant to classification was published in piecemeal fashion, scattered over the numerous technical journals serving the experts of the various fields such as:

- philosophy and science of science
- science policy and science organization
- mathematics, statistics and computer science
- library and information science
- archivistics and museology
- journalism and communication science
- industrial products and commodity science
- terminology, lexicography and linguistics

Beginning in 1974, KNOWLEDGE ORGANIZATION (formerly INTERNATIONAL CLASSIFICATION) has been serving as a common platform for the discussion of both theoretical background questions and practical application problems in many areas of concern. In each issue experts from many countries comment on questions of an adequate structuring and construction of ordering systems and on the problems of their use in opening the information contents of new literature, of data collections and survey, of tabular works and of other objects of scientific interest. Their contributions have been concerned with

1. clarifying the theoretical foundations (general ordering theory/science, theoretical bases of classification, data analysis and reduction)
2. describing practical operations connected with indexing/classification, as well as applications of classification systems and thesauri, manual and machine indexing
3. tracing the history of classification knowledge and methodology
4. discussing questions of education and training in classification
5. concerning themselves with the problems of terminology in general and with respect to special fields.

Aims

Thus, KNOWLEDGE ORGANIZATION is a forum for all those interested in the organization of knowledge on a universal or a domain-specific scale, using concept-analytical or concept-synthetic approaches, as well as quantitative and qualitative methodologies. KNOWLEDGE ORGANIZATION also addresses the intellectual and automatic compilation and use of classification systems and thesauri in all fields of knowledge, with special attention being given to the problems of terminology.

KNOWLEDGE ORGANIZATION publishes original articles, reports on conferences and similar communications, as well as book reviews, letters to the editor, and an extensive annotated bibliography of recent classification and indexing literature.

KNOWLEDGE ORGANIZATION should therefore be available at every university and research library of every country, at every information center, at colleges and schools of library and information science, in the hands of everybody interested in the fields mentioned above and thus also at every office for updating information on any topic related to the problems of order in our information-flooded times.

KNOWLEDGE ORGANIZATION was founded in 1973 by an international group of scholars with a consulting board of editors representing the world’s regions, the special classification fields, and the subject areas involved. From 1974-1980 it was published by K.G. Saur Verlag, München. Back issues of 1978-1992 are available from ERGON-Verlag, too.

As of 1989, KNOWLEDGE ORGANIZATION has become the official organ of the INTERNATIONAL SOCIETY FOR KNOWLEDGE ORGANIZATION (ISKO) and is included for every ISKO-member, personal or institutional in the membership fee (US $ 55/US $ 110).

Rates: From 2011 on for 6 issues/ann. (including indexes) € 198,00 (forwarding costs included) for the print version resp. € 229,00 for the print version plus access to the online version (PDF). Membership rates see above.

ERGON-Verlag GmbH, Keesburgstr. 11, D-97074 Würzburg; Phone: +49 (0)931 280284; FAX +49 (0)931 282872; E-mail: service@ergon-verlag.de; http://www.ergon-verlag.de

The contents of this journal are indexed and abstracted in Referativnyi Zhurnal Informatika and in the following online databases: Information Science Abstracts, INSPEC, Library and Information Science Abstracts (LISA), Library Literature, PASCAL, Sociological Abstracts, and Web Science & Social Sciences Citation Index.