Challenges in Knowledge Representation and Organization for the 21st Century. Integration of Knowledge across Boundaries

Proceedings of the Seventh International ISKO Conference
10-13 July 2002
Granada, Spain

Organized by the ISKO – Spanish Chapter and The University of Granada

Edited by MARÍA J. LÓPEZ-HUERTAS

With the assistance of FRANCISCO J. MUÑOZ-FERNÁNDEZ
Predocumentation

The volume contains:
Introduction - Keynote address - Theoretical models and universals in knowledge organization and representation - Epistemological foundations for knowledge structures and analysis - Models and methods for knowledge representation - Models and methods for knowledge organization. Tools and systems - Models and methods for knowledge organization and retrieval - Organization of integrated knowledge in the electronic environment. The internet - Models and methods for knowledge organization and conceptual relationships - Integration of knowledge in the Internet. Representing knowledge in Web sites - Models and methods for knowledge integration in information systems - Applications of Artificial Intelligence Techniques to Information Retrieval (Part I) - Integration of knowledge in multicultural domain-oriented and general systems (Part I) - Applications of Artificial Intelligence Techniques to Information Retrieval (Part II) - Epistemological approaches to classification principles, design and construction - Professional ethics. Users and information structures. Evaluation of systems - Integration of knowledge in multicultural domain-oriented and general systems (Part II) - Applications of Artificial Intelligence Techniques to Information Retrieval (Part III) - Subject Index - List of Contributors
Contents

Introduction 12

Key note address

Rebecca GREEN
Conceptual Universals in Knowledge Organization and Representation 15

1. Theoretical Models and Universals in Knowledge Organization and Representation

Jack ANDERSEN
Ascribing Cognitive Authority to Scholarly Documents on the (Possible) Role of Knowledge Organization in Scholarly Communication 28

Elin K. JACOB
Augmenting Human Capabilities: Classification as Cognitive Scaffolding 38

Clare BEGHTOL
Universal Concepts, Cultural Warrant and Cultural Hospitality 45

Maria Nélida GONZÁLEZ DE GÓMEZ
Knowledge, Communication, Information: Intersubject Links Institutional and Technological Mediations in Information 50

Joe TENNIS
Subject Ontogeny: Subject Access Through Time and the Dimensionality of Classification 54

2. Epistemological Foundations for Knowledge Structures and Analysis

Nuno SILVA and João ROCHA
Merging Ontologies Using a Bottom-Up Lexical and Structural Approach 60

Giliola NEGRINI and Patrizia ZOZI
Ontological Analysis of the Literary Work of Art 67

Jarmo SAARTI

N.Y. KOBASHI, J.W. SMIT and M. de F.G.M. TÁLAMO
Constitution of the Scientific Domain of Information Science 80

3. Models and Methods for Knowledge Representation

Anita COLEMAN
A Classification of Models 86
Gian Piero ZARRI
Indexing and Querying of Narrative Documents, a Knowledge Representation Approach

93

Jeremy J. SHAPIRO
Interdisciplinary Knowledge Integration and Intellectual Creativity

100

Rahmatollah FATTAHI and Mehri PARIROKH
Restructuring the Bibliographic Record for Better Organization, Management, and Representation of Knowledge in the Global Online Environment: a New Approach

107

Devika P. MADALLI and A.R.D. PRASSAD
Vyasa: a Knowledge Representation System for Automatic Maintenance of Analytic-Synthetic Scheme

113

Catalina NAUMIS PEÑA
Images and Words

120


Maria Inês CORDEIRO and Aida SLAVIC
Data Models for Knowledge Organization Tools: Evolution and Perspectives

Vanda BROUGHTON
Facet Analytical Theory as a Basis for Knowledge Organization Tool in a Subject Portal

135

Stella G DEXTRE CLARKE
Planning Controlled Vocabularies for the Uk Public Sector

142

Widad MUSTAFA el HADI

149

Hur-Li LEE and Allyson CARLYLE
Academic Library Gateways to Online Information: a Taxonomy of Organizational Structures

158

5. Models and Methods for Knowledge Organization and Retrieval

Gerhard J.A. RIESTHUIS and Maja ZUMER
The Functional Requirements for Bibliographic Records and Knowledge Organization

165

Rochelle KEDAR and Snunith SHOHAM
The Subject Cataloging Of Monographs With The Use Of a Thesaurus

173

Ana PÉREZ LÓPEZ, Mercedes DE LA MONEDA and Ángel MOROS RAMÍREZ
Application of the Cantor set Theory in Making Decision about the Collection Development

181

Hemalata IYER and Jeanne M. KEEFE
The WordNet as an Auxiliary Resource To Search Visual Image Database In Architecture

186
Douglas TUDHOPE, Ceri BINDING, Dorothee BLOCKS and Daniel CUNLIFFE
Representation and Retrieval in Faceted Systems

6. Organization of Integrated Knowledge in the Electronic Environment. The Internet
José Antonio SALVADOR OLIVÁN, José María ANGÓS ULLATE and María Jesús FERNÁNDEZ RUÍZ
Organization of the Information about Health Resources on the Internet
Eduardo PEIS, Antonio RUIZ, Francisco J. MUÑOZ-FERNÁNDEZ and Francisco de ALBA QUIÑONES
Practical Method to Code Archive Findings Aids in Internet
Marthinus S. VAN DER WALT
An Integrated Model For The Organization Of Electronic Information/Knowledge in Small, Medium and Micro Enterprises (Smme's) in South Africa
Ricardo EITO BRUN
Software Development and Reuse as Knowledge Management Practice
Roberto POLI
Framing Information

Terence R. SMITH, Marcia Lei ZENG, and ADEPT Knowledge Organization Team
Structured Models of Scientific Concepts for Organizing, Accessing, and Using Learning Materials
M. OUSSALAH, F. GIRET and T. KHAMMACI
A kr Multi-hierarchies/Multi-Views Model for the Development of Complex Systems
Jonathan FURNER
A Unifying Model of Document Relatedness for Hybrid Search Engines
José Manuel BARRUECO and Vicente Julián INGLADA
Reference Linking in Economics: The Citec Project
Allyson CARLYLE and Lisa M. FUSCO
Equivalence in Tillett's Bibliographic Relationships Taxonomy: a Revision
José Antonio FRÍAS and Ana Belén RÍOS HILARIO
Visibility and Invisibility of the Kindship Relationships in Bibliographic Families of the Library Catalogue
## 8. Integration of Knowledge in the Internet. Representing Knowledge in Web Sites

Houssem ASSADI and Thomas BEAUVISAGE  
A Comparative Study of Six French-Speaking Web Directories  
Barbara H. KWAŚNIK  
Commercial Web Sites and The Use of Classification Schemes: The Case of Amazon.Com  
Jorge SERRANO COBOS and Ana Mª QUINTERO ORTA  
Design, Development and Management of an Information Recovery System for an Internet Website: from Documentary Theory to Practice  
José Luis HERRERA MORILLAS and Mª del Rosario FERNÁNDEZ FALERO  
Information and Resources About Bibliographic Heritage on The Web Sites of the Spanish Universities  
J.F. ALDANA, A.C. GÓMEZ, N. MORENO, A. J. NEBRO, M.M. ROLDÁN  
Metadata Functionality for Semantic Web Integration  
Uta PRISS  
Alternatives to the "Semantic Web": Multi-Strategy Knowledge Representation  

---

## 9. Models and Methods for Knowledge Integration in Information Systems

Rebecca GREEN, Carol A. BEAN and Michèle HUDON  
Universality And Basic Level Concepts  
Grant CAMPBELL  
Chronotope And Classification: How Space-Time Configurations Affect the Gathering of Industrial Statistical Data  
Marianne LYKKE NIELSEN and Anna GJERLUF ESLAU  
Corporate Thesauri – How to Ensure Integration of Knowledge and Reflections of Diversity  
Nancy WILLIAMSON  
Knowledge Integration and Classification Schemes  
M.V. HURTADO, L. GARCÍA and J.PARETS  
Semantic Views over Heterogeneous and Distributed Data Repositories: Integration of Information System Based on Ontologies  
Fernando ELICHIRIGOITY and Cheryl KNOTT MALONE  
Representing the Global Economy: the North American Industry Classification System  

---

## 10. Applications of Artificial Intelligence Techniques to Information Retrieval (Part I)

Christopher S.G. KHOO, Karen NG and Shiyan OU  
An Exploratory Study of Human Clustering Of Web Pages
Stéphane CHAUDIRON, Majid IHADJADENE and François ROLE
Authorial Index Browsing in an Xml Digital Library 358
Xavier POLANCO
Clusters, Graphs, and Networks for Analyzing Internet-Web-Supported Communication within a Virtual Community 364
E. HERRERA-VIEDMA, O. CORDÓN, J.C. HERRERA, M. LUQUE
An IRS Based on Multi-Granular Linguistic Information 372
Pedro CUESTA, Alma M. GÓMEZ and Francisco J. RODRÍGUEZ
Using Agents for Information Retrieval 379

11. Integration of Knowledge in Multicultural Domain-Oriented and General Systems. (Part I)

Antonio GARCÍA JIMÉNEZ, Alberto DÍAZ ESTEBAN and Pablo GERVÁS
Knowledge Organization in a Multilingual System for the Personalization of Digital News Services: How to Integrate Knowledge 386
María J. LÓPEZ-HUERTAS and Mario BARITÉ
Knowledge Representation and Organization of Gender Studies on the Internet: Towards Integration 393
Victoria FRANCU
Language-Independent Structures and Multilingual Information Access 404
Annelise Mark PEJTERSEN and Hanne ALBRECHTSEN
Models for Collaborative Integration of Knowledge 412

12. Applications of Artificial Intelligence Techniques to Information Retrieval (Part II)

C. LOPEZ-PUJALTE, V.P. GUERRERO, F. de MOYA-ANEGÓN
Evaluation of the Application of Genetic Algorithms to Relevance Feedback 422
O. CORDÓN, E. HERRERA-VIEDMA, M. LUQUE, F. de MOYA-ANEGÓN and C. ZARCO
An Inductive Query by Example Technique for Extended Boolean Queries Based on Simulated Annealing-Programming 429
Víctor HERRERO-SOLANA and F. de MOYA-ANEGÓN
Graphical Table of Contents (GTOC) for Library Collections: the Application of UDC Codes for the Subject Maps 437
Luis M. CAMPOS, Juan M. FERNÁNDEZ-LUNA and Juan HUETE
Managing Documents with Bayesian Belief Networks: A Brief Survey of Applications and Models 443
13. Epistemological Approaches to Classification Principles, Design and Construction

Birger HJØRLAND
The Methodology Of Constructing Classification Schemes: A discussion of the State-of-the-Art 450
Hope OLSON, Juliet NIELSEN and Shona R. DIPPIE
Encyclopaedist Rivalry, Classificatory Commonality, Illusory Universality 457
Jian QIN
Evolving Paradigms of Knowledge Representation and Organization: A Comparative Study of Classification, XML/DTD and Ontology 465
Jens-Erik MAI
Is Classification Theory Possible? Rethinking Classification Research 472
I.C. MCILWAINE
Where Have All The Flowers Gone? An Investigation Into The Fate of Some Special Classification Schemes 479


J. Carlos FERNÁNDEZ-MOLINA and J. Augusto C. GUIMARAES
Ethical Aspects of Knowledge Organization and Representation in the Digital Environment: Their Articulation in Professional Codes of Ethics 487
Ali Asghar SHIRI, Crawford REVIE and Gobinda CHOWDHURY
Assessing the Impact of User Interaction with Thesaural Knowledge Structures: A Quantitative Analysis Framework 493
Carmen CARO CASTRO and Crispulo TRAVIESO RODRÍGUEZ
Ariadne's Thread: Knowledge Structures for Browsing in OPAC's 500
Linda BANWELL
Developing and Evaluation Framework For a Supranational Digital Library 509
Antonio L. GARCÍA GUTIÉRREZ
Knowledge Organization From a "Culture of the Border": Towards a Transcultural Ethics of Mediation 516
Christopher KING, David H. MARWICK and M. Howard WILLIAMS
The Importance of Context in Resolving of Conflicts when Sharing User Profiles 523

15. Integration of Knowledge In Multicultural Domain-Oriented and General.(Part II)

Richard P. SMIRAGLIA
Crossing Cultural Boundaries: Perspectives on the Popularity of Works 530
A. NEELAMBEGHAN and Hemalata IYER
Some Patterns of Information Presentation, Organization and Indexing for Communication Across Cultures and Faiths 539
16. Applications of Artificial Intelligence Techniques to Information Retrieval (Part III)

F. MARTÍNEZ, M.T. MARTÍN, V. M. RIVAS, M.C. DÍAZ and L.A. UREÑA

Using Neural Networks for Multiword Recognition in IR
E. PEIS, E. HERRERA-VIEDMA, J.C. HERRERA

On the Evaluation of XML Documents Using Fuzzy Linguistic Techniques
V.P. GUERRERO, C. LÓPEZ-PUJALTE, C. FABA, M.J. REYES, F. ZAPICO and F. de MOYA-ANEGÓN

Artificial Neural Networks Applied to Information Retrieval
I. BLANCO, M.J. MARTÍN-BAUTISTA, D. SÁNCHEZ, A. VILA

Fuzzy Logic for Measuring Information Retrieval Effectiveness

Subject Index

List of Contributors
Introduction

Culture, scientific evolution, and the progress of human beings, in turn, are closely bound to information as the vehicle to carry the models of knowledge and the models of the world elsewhere. For this reason, a specialty, such as ours, concerned with how to make this information successfully available to any user in any part of the world, is most challenging. At the same time, we are in a subject field that is sensible to changes of paradigms and technology.

With the turn to this new century, it is more and more obvious that we are somehow involved in a transitional process where old models are no longer standing. There are many questions but few answers as yet. This mood can be felt in different environments: Economics, Politics, Sociology, Religion, Ethics... and also in Information Science. There is a general belief that lead us to think of the universal instead of the particular. It is becoming clearer that the individual, proposed as a model of the world since the Renaissance, is no longer valid. On the other hand, technology has evolved in a way that has helped to think of universals in real time through the Internet.

We are facing times of changes. New paradigms have to respond to questions that old models cannot properly deal with. The challenging times ahead that will demand universality without exclusion, integration instead of dispersion or domination. The representation and organization of knowledge, especially in Universal Information systems, such as the Internet, need better models that allow the integration and equity of cultures and of social or numerical minorities.

With this idea in mind, the 7th ISKO International Conference has been conceived. The main objective is to address basic questions such as: 1) analysis of models for knowledge representation and organization, as a state of the art departure point and 2) the proposal of new models, methods and techniques of integrating knowledge across boundaries in order to improve performance in the new century. Theories and models for integrating knowledge from a variety of perspectives, multiculturalism and minorities, users, and information retrieval and the ethical issues that arise are the main core of the Conference.

It was interesting to see that this topic called the attention of so many scholars. The response to the call for papers was very high, and it contributed to materializing the original idea. This fact made us to decide to accept more papers than in previous Conferences, even though this means holding simultaneous sessions at certain times over the event. It is clear that thanks to the authors, their effort, work and enthusiasm, we have a Conference with plenty of interesting proposals to share. I
would say that the subjects chosen to be presented in the papers allow for a well balanced programme.

The eighty papers included in this volume refereed by an International Committee have been divided into sixteen themes: 1. Theoretical models and universals in knowledge organization and representation, 2. Epistemological foundations for knowledge structures and analysis, 3. Models and methods for knowledge representation, 4. Models and methods for knowledge organization: Tools and systems, 5. Models and methods for knowledge organization and retrieval, 6. Organization of integrated knowledge in the electronic environment. The Internet. 7. Models and methods for knowledge organization and conceptual relationships, 8. Integration of knowledge in the Internet, Representing knowledge in Web sites, 9. Models and methods for knowledge integration in information systems, 10. Applications of Artificial Intelligence Techniques to Information Retrieval (Part I), 11. Integration of knowledge in multicultural domain-oriented and general systems (Part I), 12. Applications of Artificial Intelligence Techniques to Information Retrieval (Part II), 13. Epistemological approaches to classification principles: Design and construction, 14. Professional ethics, users and information structures. Evaluation of systems, 15. Integration of knowledge in multicultural domain-oriented and general systems (Part II) and 16. Applications of Artificial Intelligence Techniques to Information Retrieval (Part III). These themes follow the same order that they have in the Conference, and the papers in this volume can be found in the same sequence.

It is interesting that many proposals deal with models for retrieval, design and contraction of conceptual structures, relations and systems and integration of knowledge. It is also important to stress that these models are presented from a wide variety of approaches, which increases the interest of the group. They are distributed in five sessions, and respond to one of the main calls of the Conference: the need for new models for knowledge integration. Another well-defined group is that related to the integration of knowledge in multicultural systems and minorities, which is also a remarkable topic in the Conference. Two sessions are dedicated to it, and very real, vivid experiences and research balance the model contribution. Integration and organization of knowledge in Internet is another important feature not only for the relevant contributions we can find in the two sessions devoted to it, but also for the additional information that can be obtained from the Open Forum on Gateways, Portals and Clearinghouses in Social Sciences. Unfortunately, only those attending the Conference can have the opportunity to access this information. Approaches to information retrieval from the artificial intelligence perspective, geared by the idea of integration, is another focus of interest. Three. Applications of Artificial Intelligence Techniques to Information Retrieval (Part II), sessions are dedicated to this topic. The light of theory upon the general theme of the Conference is an outstanding contribution here. Session one offers the opportunity to know more about universals, cultural integration, and cognitive approaches to the problem. Session two addresses the matter under more specific views. Professional Ethics conform a major issue in the Conference as can be deduced from the first paragraphs of this Introduction. It is a quite new or unusual topic in our Meetings, but of growing importance. In my view, more attention should be paid to this
matter. Nevertheless, getting started is the first step. It might be interesting to bear in mind this topic for future Conferences or Seminars. As happened in the case of the Internet, attendants to the Conference will have the opportunity to participate in a Round Table on Ethics that will complete what is included in this volume.

In summary, I would say that readers have challenging contents to explore in this book. Much thinking, concern, research and effort has come from the numerous authors that contribute to it, to whom I am most grateful. Hopefully, after four days of working and exchanging experiences in the warm summer of Granada, together we may enhance the effectiveness of information retrieval systems, challenged by a new era and new models, while heightening our sensitivity to minorities, cultural differences and ethics in a global but very compartmental world.

María J. López-Huertas
Granada, 11 may 2002
Rebecca Green  
College of Information Studies, University of Maryland, USA

Conceptual Universals in Knowledge Organization and Representation

Abstract: Within the overall conference theme—integration of knowledge across boundaries—an important subtheme is universality: Where universals of knowledge organization and representation exist, knowledge integration is more likely. Thus, knowledge of conceptual universals should inform efforts at knowledge integration. In this paper, natural language is used as a model for exploring conceptual universals, since the phenomenon of translating between languages validates, but also circumscribes, the existence of semantic and lexical universals. The paper explores a representative inventory of semantic and lexical universals that should be accounted for in knowledge organization and representation systems, especially those that aim to be comprehensive.

1. Introduction

One of the biggest challenges we face in knowledge organization and representation is heterogeneity in the expression and structure of conceptual content. We have developed a vast array of classificatory systems, for example, classification schemes, thesauri, ontologies, and lexical databases. Their number exacerbates a problem that arises naturally from their individuality, that is, (some degree of) incompatibility. Responding to this situation, we have also developed a number of tools that show correspondences across classificatory systems, some of the same type, some of different types. Examples include: conversion tables between the Library of Congress Classification and Dewey Decimal Classification schemes (Scott, 1999); correspondences within Library of Congress Subject Headings to equivalent classes within the Library of Congress Classification; cross-language equivalences built into multilingual thesauri; the mapping of dozens of medical vocabularies to the Metathesaurus® of the Unified Medical Language System® (http://www.nlm.nih.gov/research/umls/); and mappings between versions of the WordNet lexical database (e.g., http://www.cs.ucf.edu/~bmartin/wnmap/). But even where mappings across classificatory systems exist, we are not guaranteed that the end result will manifest unified semantic content and structure.

Not surprisingly, the situation that occurs with classificatory languages—on the one hand, some correspondences cut across systems, but, on the other hand, individual systems exhibit distinctive characteristics—mirrors the situation that occurs with natural languages: We can usually translate satisfactorily between them, but recognize that we have often not said exactly the same thing in doing so. On the one hand, we perceive benefit in the many natural languages and classificatory systems that exist; if but a single natural language or a single system of classification existed, we would lose the richness associated with variety; moreover, since specific language systems tend to emphasize different arenas of experience, we would probably lose the ability to reflect in depth on some aspects of our world. Multiplicity of language and classificatory systems enriches our ability to express...
conceptual content in both horizontal and vertical planes. On the other hand, if our natural languages and classificatory systems were distinctive to the point of complete dissimilarity, the existence of different natural languages and classificatory systems would stand as barriers to communication. The intellectual wealth associated with the multiplicity of language and classificatory systems would then be inaccessible to us. As a result, the situation we find with both natural languages and classificatory systems—in which many distinctive systems exist, which correspond in part, but probably only in part, to each other—is probably desirable and perhaps even optimal.

The question remains whether, or better, to what degree, it is possible to create an interlingua or switching language to which natural languages or classificatory systems could be fully mapped. Indeed, there have been numerous attempts to construct precisely such languages. It is not part of my agenda to analyze any specific such language system, but rather to explore desiderata for such tools. More specifically, my intent is to discuss some of the more important and central conceptual universals of knowledge organization and representation, with an eye to enumerating a criterial set of universal phenomena that any interlingua or switching language would have to accommodate.

2. Language Universals as a Model for Knowledge Organization and Representation

Because of the close relationship between natural languages, on the one hand, and knowledge organization and knowledge representation systems, on the other hand, a brief survey of background issues relating to natural language universals promises to illuminate our understanding of conceptual universals for knowledge organization and knowledge representation. The survey will first touch on general issues and then turn to the more specific context of translation.

2.1 Language Universals: General Background

A major landmark in the study of language universals was the 1961 Conference on Language Universals. A memorandum prepared for the conference begins, “Underlying the endless and fascinating idiosyncracies of the world’s languages there are uniformities of universal scope. Amid infinite diversity, all languages are, as it were, cut from the same pattern” (Greenberg, Osgood, & Jenkins, 1966, xv). These uniformities and common patterns, like all collections of any size or complexity, admit of different ways of being classified. One way to slice the classification pie is to consider in what stratum of language a universal operates, e.g., phonology, grammar, lexicon, and so on. Another important dimension of language universals is what Greenberg et al. call their “logical structure” (pp. xix-xxi): Do they apply to all languages (unrestricted universals) or only to some (typically, many or most) languages (statistical universals)? Do they involve a relationship between two features, such that if a language has characteristic a, it will also have characteristic b (implications), or do they apply in all circumstances ([true] universals)?

All effective knowledge organization and representation schemes impose equivalence control between various expressions of (more or less) the same conceptual content. This process reflects a concern with cognitive content on two levels. First, there is the semantic level, which addresses what counts as a conceptual unit and how conceptual units are interrelated. Second, there is the
lexical level, which addresses how conceptual content is expressed in words and phrases. Consequently, it will be language universals operating in strata related to meaning—semantic and lexical universals—that are most relevant to our inquiry. Happily for our enterprise, Immler (1991, 38) suggests that “hundreds of valid linguistic universals [could be named] as soon as we turn to the domain of semantic universals.” A significant research program has also been pursued by Wierzbicka and others (Wierzbicka, 1992; Goddard & Wierzbicka, 1994) to identify lexical primitives—concepts that cannot be defined in terms of other concepts and that appear to be expressible in all human languages; such primitives constitute lexical universals. Figure 1 summarizes the set of lexical items (as expressed in English) found to appear cross-linguistically and assumed to be universal.

<table>
<thead>
<tr>
<th>Substantives</th>
<th>I, you, someone, something, people</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mental predicates</td>
<td>think, say, know, feel, want</td>
</tr>
<tr>
<td>Determiners/quantifiers</td>
<td>this, the same, other, one, two, many, all</td>
</tr>
<tr>
<td>Actions/events</td>
<td>do, happen</td>
</tr>
<tr>
<td>Meta-predicates</td>
<td>no, if, can, like, because, very</td>
</tr>
<tr>
<td>Time/place</td>
<td>when, where, after, before, under, above</td>
</tr>
<tr>
<td>Partonomy/taxonomy</td>
<td>have parts, kind of</td>
</tr>
<tr>
<td>Evaluators/descriptors</td>
<td>good, bad, big, small</td>
</tr>
</tbody>
</table>

Figure 1. Wierzbicka’s proposed lexical and semantic primitives (Goddard, 1994, 22)

Turning to the logical structure of universals, one might initially assume that only unrestricted lexical and semantic universals—non-implicational assertions holding across all time and all languages—should matter in our pursuit of conceptual universals for knowledge organization and representation. There are several reasons to want to relax this assumption. First, the notion of unrestricted universals may be taken to presuppose a not-altogether-realistic inductive approach for postulating a universal, in which a pattern can only be claimed to be a universal if it is known to exist in all languages. But do we have full knowledge of all extant languages? What of languages that have died before being discovered by any linguist? Can we hope to predict how languages might change in the future? Second, if we somehow did have perfect knowledge of all unrestricted universals, we should presumably still want to account for the merely statistical universals—generalizations applying to many, but not all, languages—which are likely to predominate; Ullmann (1966,p. 220) claims that “most semantic universals are likely to be of the statistical variety.” Ullmann (p. 220) also suggests that “parallel developments” (e.g., certain metaphorical extensions that occur in some number of unrelated languages) warrant attention.

Immler (1991,p. 39) presents an alternative viewpoint, arguing that unrestricted semantic universals need not be based on inductive generalization: “As soon as we know the reason for the necessary presence of a feature in human language, it follows automatically that this feature must be present in all languages of the world; we then know that it must be a linguistic universal.” What would justify characterizing some conceptual content as a necessary feature of all human language? Immler takes up a variety of meanings purported to exist in all human
languages—for example, the distinctions between light and dark and between alive and dead—and then suggests “no human society could afford not to be able to express these ideas in its language” (pp. 39-40). He further argues that “cognitive contents can be universal across all cultures and all languages—they will be universal to the extent that the real world around us[—]which is the same for humans all over the world[—]is the object of cognition” (p. 49). Immler thus bases his claim for necessary semantic universals in the shared experience of human beings in the real world. Other linguists ground the universality of specific semantic components of language in human physiology, going so far as to characterize such components as “biologically given” and “innate” (Fillmore, 1971, p. 372 and Bierwisch, 1967, p. 3-4, respectively; cited in Goddard, 1994, p. 16). Another linguist maintains that basic semantic types, such as are reflected in the distinction between noun and verb, are “determined by the language-independent structure of the world (i.e., ontologically)” (Lyons, 1989, 161; cited in Goddard, 1994, p. 17).

To summarize: Investigation of semantic and lexical universals in natural languages can be expected to inform our exploration of conceptual universals for knowledge organization and representation schemes. The claim of “hundreds” of valid semantic universals and the presence of active research in the area of lexical universals lead us to expect that we may likewise posit a non-trivial set of conceptual universals for knowledge organization and representation. These universals appear to be grounded in human motivations, functions, activities, etc., stemming from our interaction with external, real world.

2.2 Language Universals: Translation

The extent to which we can satisfactorily translate between human languages is, it would seem, a good measure of the degree to which natural language incorporates semantic and lexical universals. If, on the one hand, whatever we can say in one language can be translated into (all) other languages, we might take that as prima facie evidence of an extensive array of unrestricted semantic universals; if the translation is also felicitous, we might also take that as prima facie evidence of the existence of an extensive array of lexical universals. If, on the other hand, all translations were felt to be approximations at best, there would be little hope for establishing either real semantic or real lexical universals.

Not surprisingly, there are proponents on both sides of the issue: “There are essentially two points of view from which translatability has been traditionally approached: the universalist one and the monadist one. Supporters of the former approach claim that the existence of linguistic universals ensure translatability. Those who endorse the latter approach maintain that each linguistic community interprets reality in its own particular way and this jeopardizes translatability” (Pedro, 1999). To the extent that a consensus has emerged between these positions, both extremes are rejected in favor of a compromise position in which “absolute untranslatability” is deemed not to hold and “perfect translation” is deemed “unattainable.”

Reflecting this intermediate position, Wierzbicka (1992, p. 7) notes that translation is possible, pointing to the large number of translations (“more than one thousand,” thus representing roughly one of every five or six extant languages) of (parts of) the Gospels. She also mentions, by way of contrast, the stock criticism leveled at translators: “Tradditore traditore” (‘the translator is a betrayer’), citing the view that “every language provides its own set of lexicalised concepts, . . . [and] suggests its own categorization and its own interpretation of the world—
consequently, every language is indeed a different ‘guide to reality’ (Sapir 1949,162)” (Wierzbicka, 1992, p. 20). Weaving her way between the two extremes, she argues for a somewhat limited set of lexical universals (as seen in fig. 1), which presuppose underlying semantic universals: “There are good reasons to believe that every language has words available for the basic human concepts, and that everything that can be expressed at all can be expressed by combining those basic concepts in the right way. In this sense—but only in this sense—anything that can be said in one language can be translated, without a change of meaning, into other languages” (p. 20).

Philosophical concerns are largely absent from work on machine translation. Here the assumption of basic translatability is accepted even in the face of explicit recognition of the lexical, syntactic, semantic, and pragmatic differences between languages that make the task difficult. The most established approaches to machine translation incorporate assumptions relating to lexical and semantic universals, respectively. In its most naive version, the direct method involves word-for-word translations; more sophisticated versions of the direct method include various pre- and post-processing stages, but at some point, they, too, involve identifying a lexical unit in the target language that is considered equivalent to a word or phrase found in the source language. The direct method is thus based on the general assumption that for most pairs of languages, it will be possible to identify lexical units in the two languages that are conceptually equivalent. Generalized over a large number of languages, this becomes an assumption of the existence of lexical universals across many languages. Rather than identify translational equivalents between two specific languages, the indirect method performs translation by capturing the semantic content of the source language in an interlingual representation and then generating the target language translation from that semantic representation, a more efficient approach when many languages are involved. The indirect, interlingual method is thus based on the general assumption that the conceptual content expressible in any language can be captured in some one language, although it may be artificial. This intermediate language is in essence the language of semantic universals.

In summary, although translation theorists have come down on both sides of the fence on the question whether anything expressible in one language can be expressed in every other, or even any other, language, empirical translation results suggest a more moderate stance. Lexical and semantic universals may be expected to exist, but may be somewhat limited in scope or generality. The degree to which lexical and semantic universals are constrained in human translation also limit the effectiveness of machine translation, depending on the method used.

3. Semantic and Lexical Universals for Knowledge Organization and Representation

In enumerating conceptual universals that must be accommodated by any comprehensive knowledge organization or knowledge representation system, we may productively look to the semantic and lexical universals found in natural languages. Where empirical data are available, we will adopt an inductive method, although without any serious attempt to validate a universal in all human languages; generally the validation of a pattern in some number of unrelated languages is
assumed to suffice. But we will also adopt the methodology of accepting as universals those phenomena that are grounded in shared human experience of the real world, which "no human society could afford not to be able to express."

3.1 Predicate-Argument Structure

Sasse (1991, 93) suggests that conceptual universals are more likely to exist as "universal semantic principles" than as any kind of what is here termed a lexical universal. One such principle is that of predication, which Sasse introduces in the context of the core functionality of sentences, which is to express propositions: "A proposition is normally based on the conception of an event or state, in which certain individuals may or may not be involved. . . . As far as its content is concerned, we would thus have two basic constituent parts of the sentence, the event or state . . . and the individuals . . . which take part in it. Between the [two] . . . there is a logical relation, which could be called the propositional relation, presumably identical with the concept of predication in predicate logic" (p. 77). We note that predication in first-order logic stands on the twin shoulders of predicates and the terms they take as arguments. Similarly, the data modeling theory currently in widest use recognizes two core types, entities and relationships, where relationships are defined as associations between entity classes. Relationships can readily be seen as predicates whose arguments are the entity classes they bring into association. The interaction between relationships and entities is thus analogous to the interaction between predicates and arguments. This yields three distinct types of representational systems —natural language, predicate logic, and entity-relationship modeling— all based on a similar bipartite conceptual structure, specifically, a predicate-argument structure.

Immler (1991, p. 42) notes that there are languages where it is not possible to analyze the expression of an event into predicates and arguments or where there is no single linguistic unit that performs conjunction. However, sentences from these languages can be translated into English where "the sentences as a whole are perfect (or maximal) translational equivalences one of the other" even though there is no one-to-one correspondence between the structural elements of the sentences. If predicate-argument structure is universal, it occurs on the conceptual level and is not necessarily reflected as such on the syntactic level.

However, a closely related syntactic universal has been proposed, namely, that all languages have nouns and verbs (Dixon, 1977, p. 28; cited in Thompson, 1988, p. 168). Verbs are often used to express predicates, while nouns often occur as their arguments. Not surprisingly, among Wierzbicka's proposed inventory of lexical and semantic universals (see fig. 1) are both substantives/nominals (I, you, someone, something, people) and predicates, including those expressing actions/events (do, happen) and those expressing mental predication (think, say, know, feel, want) (Goddard, 1994, p. 22). Thus, even if the universality of predicate-argument structure cannot be supported strictly on syntactic grounds, still it is so central to our conceptual systems that in many, if not in most, cases the verb/noun distinction will reflect a corresponding predicate/argument distinction. Indeed, not only might we assume that all languages have some means for expressing predicate-argument structure, but we might also assume that almost all language use attempts to express predicate-argument structure that occurs at the conceptual level. Accordingly, our investigation of further universals will be guided by the predicate-argument distinction.
3.2 Characteristics of Arguments: Individuation, Classification, and Countability

We tend to think of that which exists in the real world as entities or things, typically corresponding to the argument side of the predicate-argument distinction. To the extent that the real/physical world exists independently of or logically prior to human cognition and to the extent that such real phenomena exist throughout the world, they are likely to occur as lexical universals—not necessarily in the narrow sense of being lexical primitives, but in the broader sense of being expressible (and probably lexicalized) in most/all languages. Immler (1991, p. 39, quoting Immler, 1974, p. 41) proposes that “in every language of the world there are names, designations for the following objects...: ‘soil, [many animals], [many plants], [parts of the body],... air, water, rain/snow, wind, sun.’” In like manner, Antal (1963, p. 85; quoted in Immler, 1991, p. 49) suggests: “Although the world of meanings is different and theoretically incomparable and incommensurable from one nation to another, the denotata amongst which people live are, in fact, the same for all humanity or, at least, for a great number of people.”

The previous claim that “to the extent that... real phenomena exist throughout the world, they are likely to occur as lexical universals” is on some level naive, since the lexicalization process is driven in large part by human cognition/experience. Thus, the omnipresence of sun, moon, stars, and sky for all humans does not lead necessarily to a shared conceptualization of these phenomena. One can imagine the (at least logical) possibility of seeing them as fewer than, or more than, four types of things. And yet one can also imagine that most cultures will see them as exactly four types of things.

At issue here is, as a first step, the universality of what constitutes a single entity and, as a second step, the universality of which entities are seen as the same type of thing, i.e., belong to the same class. Full universality at the second stage is severely constrained not only by the partial nature of the distribution of things throughout the world (for example, some climates never experience frost or snow; flora and fauna may be confined to certain regions of the world), but also (and as importantly) by cultural and individual differences in knowledge, values, resources, etc. Still, we can expect where phenomena exhibit salient and discrete characteristics (of, for example, shape and color) and are of a size to make them easily observable by humans, universality of perception of what is a single entity unit is the more likely. As there are multiple such entity units which differ from each other in small ways, but which are distinct from other entity units (i.e., differ from them in significant ways), universality of perception of what belongs in the same class is also the more likely. Some things and groups of things in our world fit this description, but many others do not. Hence we can expect universality in these arenas to exist, but only in limited areas. Universality of class membership is of necessity constrained by universality in unit perception.

Although the universality of unit perception and the universality of class membership operate in only restricted arenas, the underlying phenomena—individuation and classification—are universal; and dependent on both individuation and classification we have another universal, that of countability.

Individuation is the principle of picking out specific regions within our world(s) and endowing them with independent ontological status, the principle of particularization or identification. Built into the principle of individuation is a basic distinction between what is the same thing and what is a different thing. Although
there are many situations in which we need to deal on the level of particulars—one should be careful, for example, not to drive away in any vehicle other than the specific one to which one holds the title, no matter how similar they may be—there are many other situations (probably far many more) in which individuation is neither necessary nor desirable. Any appropriately valued token or set of tokens may be used in a telephone booth, for example. It is not simply for purposes of cognitive economy that we so often deal at the level of classes rather than at the level of individual entities; much of our reasoning is done at the class level (Sokal, 1974). Thus we would expect the distinction between classes of entities and individual entities, that is, the type/token distinction, to be universal.

Once we recognize both individual entities and classes comprised of multiple entities of the same kind, we come quickly to the notion of quantifying the membership of the class: How many individual entities are perceived as being of a kind? This quantification is effected through counting individual entities. A related process of quantification can be applied to entity regions that do not undergo individuation; these are perceived in terms of mass quantity, not count quantity. In this case units of measure (e.g., liters, cups, handfuls) substitute for individual entities as countable items, individuation of some sort being a prerequisite to counting (Link, 1991, p. 133). In either case, whether we perceive the entity being measured in terms of mass or count quantification, countability produces a universal set of integers. Another extension of countability gives us the difference between ‘some’ and ‘all’, the distinction between existential and universal quantification.

A personal application of individuation generates a sense of self that is distinct from everything else in all ‘normal/healthy’ persons (this sense of self would generally be considered criterial for normalcy and health). Consequently, all languages can be expected to have some notion equivalent to first person singular. The application of classification to each *I* results in first person plural, *we*, which is likely to admit of different memberships in different contexts. Meanwhile, all persons not part of *us* are, collectively, *them*, third person plural, or, individually, *he/she/it* (the gender breakdown for third person not being established as a universal by anything examined thus far; similarly, the recognition of second person arises from the phenomenon of communication and is thus less philosophically central than first and third persons).

In summary, for substantives we accept the full universality of several basic principles—individuation, classification, and countability. Furthermore, we assume the universality of some limited number of real world phenomena.

### 3.3 Characteristics of Predicates: States, Events, and Roles

On the predicate side of the predicate-argument distinction, we have a potentially large set of conceptual relationship types, typically broken down into states and events, with states as static relationships and events as dynamic relationships (although the breakdown is not nearly so clean as suggested: where, for example, do such cognitive predicates as believing, remembering, and recognizing belong?)

We discussed previously the real world grounding of some entities that are likely to be universally recognized, for example, the sun. Such universals are based on the omnipresence of certain substantives in the experience of essentially all persons. A parallel argument can be made for the universal recognition of various physical states that exist in nature and that are likely to impinge on the life of all
persons, for example, light vs. dark, wet vs. dry, warm vs. cold, alive vs. dead, big/long/wide/deep vs. little/short/narrow/shallow, hard vs. soft, heavy vs. light, smooth vs. rough, young vs. old, sick vs. healthy, pain vs. pleasure (Immler, 1991, p. 39 [also citing Immler, 1974, p. 41]; Thompson, 1988, p. 168); the latter few states are more likely to be known from personal experience, although many persons would also recognize them from animal life around them.

Other stative universals seem intrinsic to human physiological experience. These include such states as hunger, thirst, weariness, and a specific set of emotions (joy, sorrow, fear, disgust, anger, surprise) whose facial expressions are recognized cross-culturally (Ekman, 1993). Similarly, although color boundaries are language-specific, the perception of focal colors is apparently universal (Kay, 1999).

Additional states of a universal nature arise from the human social experience. Such might include basic kinship relationships (at least the general notion of kinship relationships) and evaluative properties (good vs. bad).

Dynamic universals would probably similarly arise from real world phenomena (e.g., weather events, manners of locomotion—running, flying, swimming—observed in animals), from human physiology (e.g., eating, drinking, sleeping; vision, touch, taste), and the basic social, cognitive, and psychological nature of humans (e.g., verbal communication, birth, belief/thought/memory, desire).

Predicates are often characterized in terms of the semantic roles of their arguments, where a semantic role is a functional semantic type at the most abstract level. Attempts have been made to enumerate a universal set of semantic roles, but the endeavor is probably ill-fated. Davis (1994, p. 161, 163) suggests that recognizing their “constant presence in language” is a more productive approach to semantic roles than trying to develop a universal inventory.

However, as we have already repeatedly seen, not being able to corroborate that a comprehensive set of universals of some sort exists is not tantamount to denying that any universals of that sort exist. For example, Comrie (1989, p. 59) notes with respect to the roles of agent, force, instrument, experiencer, and patient that they are “not so much a set of discrete semantic relations, but rather a continuum . . . of control.” These roles hold central positions in those semantic role inventories that have been proposed. We may therefore assume that, even though they may not constitute discrete cross-linguistic semantic roles, control is probably a central notion in all languages. If so, we may also assume the universal centrality of animacy, which correlates closely with control.

To the extent that roles can be identified with general universal predicates—for example, event or state, at the most general level, or communication or locomotion, at a somewhat more specific level—we may likewise find that those semantic roles are universal. Here we take the example of events, which take place in time and space. Given the linguistic universality of events, we would expect to find time and space as semantic roles in all languages. In a cross-cultural study of temporal collocations and metaphors in English, Mandarin, Hindi, and Sesotho, Alverson (1994) goes further, claiming that time is experienced and expressed in similar ways universally: “as a partible entity, . . . as a causal agent with effects, . . . as a medium in motion, . . . as a linear and as a circular course, . . . and . . . as its method of ascertainment” (p. 104). Likewise, although spatial expressions in different languages display syntactic variation, there are underlying patterns that cut across them: objects (trajectors) are located in relation to other objects (landmarks),
in terms of either the coincidence or separation of their locations, construed in 1-, 2-, or 3-dimensions (Hawkins, 1993).

In summary, for predicates we accept the full universality of some specific states and events—probably numbering in the dozens—that occur in the world and/or that arise from common human experience. Further, we recognize the universality of semantic roles; although it may be impossible to enumerate a well-defined and restricted set of roles, still some notions underlying semantic roles appear to be universal: control, as a universal parameter of roles, on the one hand, and time and space, as universal roles, on the other hand.

3.4 Logical Operators

So far, we have been operating in the context of a bipartite conceptual structure based on predicates and their arguments. Weinreich (1966, p. 148) takes this line of reasoning further when he claims, “It will be assumed that it is possible to describe all discourse as either (a) having the semiotic form ‘$\text{Of}(x)$’, or (b) deviating from it in specified ways. In this formulation ‘x’ stands for an argument—‘something talked about’; ‘f’ for a predicate—‘something said about x’; and ‘$\text{O}$’ is a covering label for any of a number of operations.” Weinreich therefore proposes the universal existence of a set of operators, which supplement our fundamental categories of predicates and arguments. Some of Weinreich’s operations are strictly linguistic in nature and of no particular consequence to us here. Other notions that Weinreich treats in terms of operators (e.g., person, time, place, same vs. different, count vs. mass), we have handled elsewhere. There remain the propositional operations of negation, disjunction, conjunction, implication, and equivalence. Weinreich acknowledges that “no language represents such operations with the maximum economy” (p. 158), meaning that a one-to-one correspondence between linguistic expressions and logical operations is often lacking. Indeed some amount of circumlocution may be required to express such concepts; moreover, the means available for expressing them may include as well the expression of other semantic components. That it is logically possible to define all these operations in terms of negation and conjunction further obscures the picture. Nevertheless, we can propose reasons why most of these logical operations would be universal. Negation is implied by antonymy, which is considered a universal (see section 3.5). According to Weinreich, linking, an effect of conjunction, is one of two semiotic processes (along with nesting) by which linguistic signs can be “grammatically combined to form discourse” (pp. 163-164). If, as he claims, “in all languages a combination of signs takes the form of either linking or nesting, and all languages use both patterns in kernel sentences” (p. 167), then conjunction, too, must be universal. Disjunction (with its intimate relationship to choice) and implication (with its close connection to causation) probably are also universal.

3.5 Lexical Relationships

The standard paradigmatic/lexicosemantic relationships which must be handled in thesauri—taxonomy, partonomy, synonymy, antonymy—are likely to be found in all languages. This generalization implies that all natural lexicons are structured (at least eventually so) and that human-constructed languages are simply mirroring relationships that occur in natural language.

Evidence from empirical studies of many languages supports the universality of such hierarchical relationships as the subsumption/taxonomy/IS-A relationship and the part-whole relationship (Goddard & Wierzbicka, 1994). In a similar vein Ullmann (1966, p. 231) refers to cross-linguistic evidence for the universality of
synonymy wherever there is, as there is always likely to be, specialization: “It has often been found that subjects prominent in the interests and activities of a community tend to attract a large number of synonyms.” Similarly, the presence of antonyms cross-linguistically argues for the admission of antonymy as a language universal (Lehrer & Lehrer, 1982; Raybeck & Herrmann, 1996).

Ullmann argues for the universality of polysemy on the basis of both cognitive economy and the richness of language: “Polysemy is in all probability a semantic universal inherent in the fundamental structure of language. The alternative to it is quite unthinkable: it would mean that we would have to store in our brains a tremendous stock of words, with separate names for any possible subject we might wish to talk about; it would also mean that there would be no metaphors and that language would thus be robbed of much of its expressiveness and flexibility” (p. 232). But, according to Greenberg et al. (1966, p. xxii), “All languages have some metaphorically transferred meanings.”

4. Summary and Conclusion

What are the implications of this survey of semantic and lexical universals for the existence of conceptual universals in knowledge organization and knowledge representation systems and (2) for knowledge integration? First, we are led to reject two extremes. On the one hand, we cannot reasonably hope to build a super-system (whether it be a classification scheme, a thesaurus, an ontology) that faithfully integrates all extant instances of that type of scheme. Individual schemes may be unique in the set of substantives and predicates they recognize as well as how they relate conceptual units to each other. On the other hand, we also cannot responsibly throw up our hands, supposing we cannot do anything. There appear to be (dozens of) conceptual units (e.g., sun, bad, think) that are more or less universally apprehended. There are also some number of conceptual abstractions (e.g., predicate-argument structure, countability, animacy, hierarchy) that are used universally in structuring our conceptual systems. Any two knowledge-based schemes are likely to be alike in fundamental ways.

How can we build upon that which knowledge organization or knowledge representation schemes share with each other? Obviously, where specific conceptual units are shared, we should construct cross-system mappings; another paper at this conference (Green, Bean, & Hudon, this volume) reports preliminary data suggesting that this sharing is more likely to occur at the basic level than at hierarchical levels above or below the basic level. It may be that the best we can do in integrating two knowledge-based schemes is to construct such crosswalks as are truly appropriate and then rely on the individual structures of the schemes being “integrated.” (Indeed, this is the approach of several major efforts of this type, e.g., EuroWordNet and the Mctathesaurus® of the Unified Medical Language System ®.)

As for the shared conceptual abstractions, they may be best used as parameters for analyzing any two systems that are subject to integration efforts. For example, what differences exist between the set of substantives treated as count nouns vs. those treated as mass nouns? Where do the semantic roles implied by the two systems fall on a control continuum? There is no magic key to knowledge
integration. However, knowledge of these conceptual universals can at least facilitate and illuminate the very considerable work that knowledge integration will always require.

Whether we can do more will be subject to continued exploration, but this much we can do: (1) Recognize true conceptual equivalences across systems; these are most likely to be found in those areas of knowledge-based systems that address real world phenomena at the basic level. (2) Retain the structural integrity of individual systems. (3) Use universal conceptual abstractions to analyze these systems, so as to further our understanding of the nature of their differences.

Notes
1. Some of the more recent approaches taken to machine translation (example-based translation, statistical translation) do not engage in isolating the meaning of a text in the source language before generating a meaning-equivalent expression in the target language; instead they depend on the pre-existence and availability of a large parallel corpus, in which extensive amounts of text have already been translated and in which corresponding units of source and target texts are aligned with each other. Since no meaningful semantic representation is established, these translation models do not further our understanding of semantic universals.

References


Jack Andersen
Department of Information Studies, Royal School of Library and Information Science, Copenhagen, Denmark.

Ascribing Cognitive Authority to Scholarly Documents. On the (Possible) Role of Knowledge Organization in Scholarly Communication.

Abstract: The new electronic environments pose a threat and challenge to the theory and practice of knowledge organization. Documents can be approached in electronic retrieval activities in ways not dependent on 'classical' knowledge organization activities such as indexing or classification. Accordingly, an argument stating the qualitative difference of knowledge organization in the new electronic environments must show that knowledge organization is worth pursuing and that it is a valuable support to users of information retrieval (IR) systems. In this paper the qualitative difference of knowledge organization and its role in scholarly communication is framed as a question of ascribing cognitive authority to documents. The concern is to examine and discuss how and to what extent knowledge organization as an epistemic instrument in scholarly communication can contribute to ascribe cognitive authority to scholarly documents. The paper is structured in the following way. Initially, a brief examination of the appearance of cognitive authority in knowledge organization, and how that affects an argument stating the qualitative difference of knowledge organization shall be presented. Secondly, the theoretical approach will be outlined and discussed. Then the empirical analysis applying the theory will be presented. The last part will point to the benefits, limitations, and possibilities of the proposed theoretical approach in relation to the conception of knowledge organization as an epistemic activity in scholarly communication.

1. The Argument: Cognitive authority in knowledge organization

When organizing and representing documents in information systems for the purpose of retrieval, these documents are the product of an already established social organization of knowledge or division of labor. They are so because they have been produced with aim of fulfilling or achieving some human activities organized socially. Insofar the social organization of knowledge expresses a kind of cognitive authority this also means that there are several levels of cognitive authority present in a document representation. A traditional document representation consists of various textual elements such as author, title, publisher, institution (or corporate source), journal name, abstract, index terms, and, in the case of the citation indexes, the reference list. It can be argued that these textual elements both in themselves and together contribute to the cognitive authority of a document. These levels can all be seen as communicative interactions mediating between document, author, publisher, library, and reader. Moreover, the scholarly journals are but one expression of the social organization of knowledge within a scholarly domain. They are (usually) organized around some particular subject matter. Author and reader are also organized within the social activities of writing and reading that tie them to the social organization of the domain. The publishers are also part of the
social organization of knowledge. They are responsible for the distribution of the material output (i.e. the documents) of the disciplines both to the esoteric circles of the disciplines and a wider public. Thus, the knowledge materialized in scholarly documents, and the object for knowledge organization, is the product of already socially organized human communicative activities structured around the social interactions between author, document, reader, and publisher. Bazerman (1988) illuminated this in his analysis of the experimental article in science. He paid particular attention to how this genre emerged historically and how it was, and is, shaped by the typified communicative activities of writers, editors, and readers, and how the genre shaped the knowledge producing activity into a typified activity exactly as a product of history. In addition the actual scholarly document itself may also have an inherent cognitive authority in terms of the credibility, validity, and reliability of what it reports, its theoretical claims, methodology, arguments, or the validity of the interpretation given (Hjørland, 1992a). Thus, given there are already several levels of cognitive authorities present in and ascribed to a document in advance, what is the role left for knowledge organization? This is a theoretical question facing knowledge organization and it must try to offer an answer. Knowledge organization must seek to spell out how it can contribute to the overall characteristic of documents in the communication of knowledge in scholarly communication. This can, among other things, be framed as a question of cognitive authority (Wilson, 1983). For that reason, knowledge organization, as an epistemic instrument in the control of writings (and reading), should help scholarly communication work optimally, and it must try to explain how and to what extent it can contribute to ascribe to documents cognitive authority. By epistemic instrument is meant knowledge organization as a means to maximize over time the intellectual utilization of recorded scholarly knowledge and minimize falsity, error, or ignorance in this process. This is interpreted as being in accordance with Wilson’s (1968) notion of “the best textual means to an end” in connection with bibliographical control, and Hjørland’s (1992b, 1997) “epistemological potentials of documents” in connection with subject analysis.

2. Theoretical framework: A Genre and Writing Approach

In an article analyzing the trajectory of scientific facts from primary literature to popular literature, Fahnestock (1986) presented this as both a shift in rhetorical situation and in genre. With this, Fahnestock points to that because of this, scientific facts cannot just be distilled without a loss of certainty and credibility. In a similar way how can the shift in rhetorical situation and genre be accounted for when it comes to the trajectory of knowledge claims from primary literature to bibliographic literature without a complete loss of certainty and credibility? Being two different genres, primary literature and bibliographic literature have different rhetorical problems because they constitute two different, though interdependent, activities in scholarly communication. The former has to produce, or make claims for, new knowledge. The latter registers and describes primary literature for the purpose of retrieval and bibliographic control. It seems to be essential for the functioning of both primary and bibliographic literatures in scholarly communication that this change is accounted for and understood in order for the bibliographic literature to be an adequate epistemic response to the knowledge producing activity. Studies of scholarly writing (e.g. Bizzell, 1982; Bazerman, 1988; Myers, 1990; Swales, 1990;
Selzer, 1993; Prior, 1998), emphasizing it as a situated and mediated epistemic activity embedded in various discourse communities, have proven to be a fruitful way of understanding the knowledge producing activities, scholarly literature both shapes and is being shaped by. By treating scholarly writing as a product of situated socio-historical and literate activity, and by relying on Miller’s concept of genre as 'typified rhetorical actions based in recurrent situations' (Miller, 1994), these studies have pointed to how any claim for knowledge is embedded in typified rhetorical activities. Hence, by uncovering the rhetoric, an insight into what constitutes scholarly knowledge and how it helps organize the recurrent situated knowledge producing activities can be gained. Given that the role of knowledge organization in written scholarly communication can be seen as that of providing and optimizing intellectual access to scholarly literature, it seems reasonable to assume that this cannot be achieved without recognizing the activity of scholarly writing. Thus, it is argued in this paper that an examination of the way scholarly documents are written can contribute to an understanding of its knowledge producing activity and to a more convenient or appropriate organization and representation of scholarly documents in information systems. Following Bazerman (1988), I will analyze a primary article in relation to four contexts as to how they are referred to, invoked, or acted on in the article. The four contexts are 1) the object under study; 2) the literature of the field; 3) the anticipated audience and 4) the author’s own self (Bazerman, 1988, p. 24). Extending this analysis to also cover the corresponding organization of knowledge claims in knowledge organization systems (e.g. bibliographies, catalogs, or classification systems), the objective is to examine the relationship between the knowledge the documents claim to produce and contribute with, and how they are in fact indexed. This can account for the shift in genre that takes place when indexing scholarly documents in information systems.

3. Empirical study: The Case of a LIS-article

The article chosen for examination is Harter's (1992) article on psychological relevance. In here Harter sets out to introduce Sperber & Wilson’s psychological concept of relevance presented in their book ‘Relevance: Communication and Cognition’ (1986). Psychological relevance is what affects or causes a cognitive change in the mind of the user. Harter’s agenda implies two things. First, Harter has to believe that this kind of relevance is useful for LIS. Second, he has to show this usability, and thereby argue in favor of it. Consequently, Harter finds himself in an epistemic-rhetorical situation. The significance of this epistemic-rhetorical situation is further underscored by the fact that relevance research since the Cranfield experiments has been, and still is, a classic area of research in information retrieval (IR), challenging Harter with a heritage that he must take into account. Moreover, Harter's article is a theoretical article about a concept that is born out of a highly experimental tradition. Thus, Harter speaks through a particular research genre. What, then, does Harter contribute with? The object under study in Harter’s article is the concept of relevance. Harter begins his article with contextualizing the historical rootedness of relevance research in LIS. In the opening paragraph of the Introduction relevance is referred to as ‘central to information retrieval (IR)’ and ‘as the basis of nearly all experimental evaluation and testing.’ (p. 602). As a research area relevance is further invoked, when Harter identifies two classic relevance ‘camps’ in IR research: The objective sense of relevance, based on the Cranfield
model, and the subjective sense of relevance. The former contemplated in terms of the topicality of the search request and document(s) retrieved and the degree of match between these two. The latter as the view that takes the relevance judgments of the user into account. That way Harter establishes a context for his upcoming argument and analysis. Harter invokes the literature of the field in two ways. When talking about the objective view of relevance the literature invoked is the critics of this view, no proponents are cited. In continuation of this, Harter comments that despite the critique of the Cranfield model ‘...the model has been accepted, generally uncritically, as the basis for most experimental research done on IR systems to the present day.’ (p. 602). Thus, the author’s own self become present when commenting on the acceptability of the Cranfield model and on the lack of critical distance to it. Harter’s critical attitude toward objective relevance is further underlined when stating that: ‘The idea that users want, or will be most happy with, documents on the topic of a search statement is an assertion that has absolutely no empirical support.’ (p. 603). Thus, relevance is also referred to as an empirical matter. As a consequence of this, Harter makes himself present by claiming that ‘I believe that a view of relevance that equates relevant documents with documents “on the subject” has been pernicious, both to the ultimate goals of the user and to the development of theory in information science.’ (p. 603). Thus, Harter is very clearly distancing himself from objective relevance. He does by making certain rhetorical moves: He does not cite proponents of the view. This can be explained by the fact that it may be assumed to be known by the audience. He emphasizes the critics of the Cranfield model. He further invokes his audience by claiming that objective relevance does not have empirical support, implying that if it had, it could have been more acceptable. Further, objective relevance is accused of having been a major hindrance for theory development in LIS. This is a strong attack on objective relevance implying that Harter’s contribution must be considered as something that can contribute to and re-create theory development in LIS. In contrast to objective relevance, proponents of subjective relevance are cited by providing what Swales (1990, p. 148) calls non-integral citations. That is, citations occurring in parentheses rather than part of the citing sentence as a sentence element. Different notions of subjective relevance such as pertinence, situational relevance, perceived utility, informativeness, beneficiality are referred to. Furthermore, the review of relevance given by Schamber, Eisenberg & Nilan (1990), arguing for a dynamic and user-oriented view of relevance, is characterized as excellent. Without having introduced readers yet to the ideas of Sperber and Wilson (1986), Harter nonetheless connects the psychological concept of relevance to the subjective view of relevance in LIS: ‘It roughly incorporates or is consistent with the notion of pertinence, situational relevance, perceived utility, informativeness, beneficiality, and other types of subjective relevance... However, without a supporting theory, subjective relevance is too vague to offer much insight.’ (p. 603). Hence, Harter claims that subjective relevance needs a theory and offers Sperber & Wilson’s theory of relevance as a solution: ‘Not only is Sperber and Wilson’s theory of psychological relevance directly applicable to the notion of relevance in information retrieval, it offers valuable insights into its essential nature.’ (p. 604). Thus, already in the Introduction the theory is decided to be applicable and valuable for IR. Accordingly, in the Introduction Harter has set the stage for his further knowledge claims. He has dissociated himself with the notion of objective relevance and sets out to argue that Sperber & Wilson’s psychological relevance can be a supporting
theory for subjective relevance. The latter needs that and the former has been a hindrance for the development of theory. That way, the concept of relevance (the object of study) is given an epistemic-rhetorical context from where to argue, discuss, and analyze it. Sections three and four of the article, 'Psychological Relevance and Information Retrieval' and 'An Extended Example' are decisive for Harter's agenda in that he has to show the applicability of Sperber & Wilson's relevance theory to IR; i.e. his contribution to knowledge is what is going to be presented here. This is done through theoretical analysis and exemplification. Because Harter is about to make a knowledge claim, it is important for him that he invokes his audience since he is not allowed to state whatever he likes. In view of that, the audience is invoked as being part of Harter's activity of interpreting the theory: 'Let us see how psychological relevance can be interpreted in terms of a document retrieval system.' (p. 606). Thus, the interpretation of the theory is a collective activity meaning all participants take responsibility. This way of invoking the audience is repeated when revising the definition of information need on the basis of the theory of psychological relevance: 'Let us revise our definition of information need, then,...' (p. 606). Having analyzed psychological relevance in relation to IR, Harter once again invokes his audience, signaling the collective nature of the activity: 'Psychological relevance allows us to talk about an "information need" as the current context -- the cognitive state at a given time -- of an individual who consults an information system.' (p. 607). Accordingly, the object of study (i.e. relevance) is turned into a psychological issue. However, when finishing this section, Harter makes clear that he has done the interpretation but with permission from his audience: 'My interpretation of information need suggests that we will...' (p. 607). Thus, when Harter has to argue for his contribution to knowledge and show its innovativeness, he invokes his audience in this activity and re-interprets old conceptions of information need in light of psychological relevance. The explanation for this may be due to that knowledge claims, following Fleck (1979), in journal articles are at the time of publication not yet to be considered a part of the certified body of knowledge of a research community (the thought collective in Fleck's terms). The topics or problems treated in a particular journal article must therefore in some way or another address topics or problems already known to the thought collective in order to acquire a possible acceptance. This is in line with Fleck (1979, p. 98) when he defines a fact as "...a stylized signal of resistance in thinking." and because the thought style is determined by the thought collective, a fact is then designated "...as the signal of resistance by the thought collective." (Fleck, 1979, p. 98). Harter exactly addresses psychological relevance in relation to topics already known by the LIS research community in order to overcome, or deal with, the signal of resistance, and to turn his interpretation of psychological relevance into a signal of resistance, into a fact. Otherwise he cannot claim to have made a contribution to knowledge. After this theoretical analysis and the establishment of his contribution to knowledge, Harter goes on with exemplifying his new theory, as a way of demonstrating to his audience the concept of psychological relevance and how it can be interpreted in relation to LIS-issues. That way he invokes a familiar ground for both author and audience. Turning to the implications of the concept of psychological relevance for IR, Harter has to demonstrate how his contribution to knowledge makes a difference. He does this by discussing the concept in relation to a range of IR research issues. Throughout this section the literature of the field is invoked when
something similar to Harter's suggestion has been proposed or when the critics of objective relevance are referred to. Like in the Introduction, potential proponents of the objective view of relevance are not cited. It is assumed that this is part of the common knowledge of the audience. This is clearly illustrated when discussing the notion of information need and how it has been employed in IR: ‘...it [information need] has usually been taken as static – representing something that is fixed and unchanging – even in very recent writings; the work of Schamber, Eisenberg & Nilon (1990) and Katzer and Snyder (1990) are rare exceptions.’ (p. 610). The ‘very recent writings’ are not given examples of, while those representing an exception are cited. Harter thereby invokes his audience through his passive and active use of the literature. Showing the implications of psychological relevance to selected information seeking studies, Harter makes his presence and refers to psychological relevance as has been ever present in these, but not explicated: ‘I suggest that psychological relevance may be the concept that underlies these models of the information-seeking process, and that unites them. Psychological relevance helps explain why these models work, as behavioral descriptions of information-seeking behavior.’ (p. 611). By arguing that psychological relevance underlies and can help explain the successes of these models, Harter seeks to justify the concept and thereby shows its role as a supporting theory. Those who do not agree with this then, now have the burden of proof for showing this is not so. Because psychological relevance implies, in Harter’s interpretation of it, changing and dynamic relevance judgments, retrieval tests and relevance judgments based on the measures of recall and precision are referred to as wrong (p. 612). Harter further underscores this disapproval of retrieval tests by making one of the most explicit claims and evidence of his persona in the article: ‘I no longer believe that there is a valid interpretation of the meaning of the results of such tests.’ (p. 612). By referring to the results of traditional retrieval testing as invalid and wrong, Harter has just refuted four decades of traditional, experimental IR research. This is a strong claim, which presumably cannot go unnoticed and uncommented in the IR research community. It could be speculated that this claim has been the motivation for the writing of the article. Throughout the entire article the concept of psychological relevance is constantly held in opposition to this established research area. Further, the sentence ends the paragraph and no other evidence is invoked or referred to in order to strengthen the claim. It is a statement that needs no further explanation, insofar the theory of psychological relevance is accepted. When analyzing the implications of psychological relevance in connection with IR and bibliometrics, relevance is invoked as the notion that brings these two research areas together (p. 612). Psychological relevance and its relation with citation practices is invoked as a measurable and empirical concept that needs to be tested in order to ‘prove’ its applicability. Thus, besides the theoretical approach taken in the article, Harter also appeals to and recognizes the need for empirical research on this matter. An article co-authored by Harter is here referred to as an example of an empirical test that has been conducted. In the last section of the article suggestions for further research are proposed. Once again psychological relevance is invoked as something that in future research needs to be empirically tested and measured (p. 614). Harter poses several questions as to how this can be done and what it will imply. This is also a way of accommodating his audience. Harter has not presented empirical evidence in his article, but acknowledges that psychological relevance needs an empirical basis if it is going to have a future role to play in LIS research.
Thereby, Harter accommodates the kind of audience that wants empirical (experimental) evidence in favor of a theory. The audience is invoked several times by the use of ‘we’, ‘us’ and ‘our’ when Harter is pointing to the need for investigating, or re-investigating, LIS research topics in light of psychological relevance. Once again it is being stressed that psychological relevance needs to be tested empirically (p. 614). Consequently, the structure of the article can be regarded as a rhetorical structure because it guides Harter’s overall argument and his knowledge claims. The relationships between author, audience, literature and the object of study are in constant interaction, demonstrating how knowledge production can be seen as a situated literate activity. How then is Harter’s article indexed in LIS-bibliographies? How is this activity accounted for? LISA has assigned to the article the descriptor-string “Information science; Implications for; Psychological relevance”. Where does such kind of indexing lead to? What kind of activity does it serve? The implications of psychological relevance for information science is also part of Harter’s knowledge claim. If it is accepted as a suitable relevance concept for LIS, then it has some implications for existing LIS research and then the indexing to some extent reflects the knowledge producing activity of the article and contributes to its cognitive authority. However, the title of Harter’s article, ‘Psychological Relevance and Information Science’ indicates that this article is about the implications of psychological relevance for LIS. Further, if title access does not fully suggest this, abstract access certainly does. Here it is stated that the implications of psychological relevance for LIS is treated in the article. On that point it could be argued that the indexing undertaken is redundant because it does not make a qualitative difference, though pointing to Harter’s knowledge claim. In Information Science Abstracts the article is indexed with these descriptors: “Bibliometrics”, “Information retrieval”, “Information science”, and “Psychological aspects”. These descriptors are broad and do not as such refer to relevance. “Bibliometrics” is not a totally odd descriptor. Harter suggests that psychological relevance underlies the citation behavior of scholars (and his also suggests that his article is not about bibliometrics (Harter, 1992, p. 603)) and that IR and bibliometrics are connected through this notion of relevance. This is part of his general knowledge claim. The descriptors “Information Retrieval” and “Information science” must be considered inappropriate. As a descriptor “Information Retrieval” does not reflect, or relate to, Harter’s overall knowledge claim or argument, but is of course affected insofar his knowledge claim is accepted and as an access point it does not have enough discriminating value. This also applies to the descriptor “Information Science”. In a bibliography covering LIS, the descriptor is useless since all documents in terms of their knowledge claims must necessarily either be about or relate to LIS. Indeed it must be the criterion for inclusion in the bibliography. From a retrieval point of view, the title provides access through this term suggesting that also indexing the document with that descriptor is redundant. Despite its discriminating value, “Psychological aspects” is ambiguous as a descriptor. Harter clearly implies psychological aspects with his knowledge claims. However, psychological aspects in the present of Harter’s article are not specific enough since this does not indicate what kinds of psychological aspects (e.g. cognitive, behavioral, or cultural) there are at stake. The indexing of Harter’s article in two LIS bibliographies reveals that it does not fully reflect the knowledge producing activity of Harter’s article. The shift in genre implies that Harter’s knowledge claims are not clearly accounted for and the indexing
undertaken cannot be justified in terms of its contribution to the cognitive authority of the article. What, then, can the above close reading of Harter’s article contribute with in terms of appropriate descriptors that reflect, or mediate, his arguments and knowledge claims? Harter examines Sperber & Wilson’s (1986) concept of psychological relevance and proposes it as an appropriate relevance concept for LIS studies. It implies re-thinking the whole notion of relevance within LIS. To do that Harter carries out a concept study with the aim of arguing for and showing its relevance for LIS. In this sense a descriptor reflecting his mode of argument and his knowledge producing activity could be “Concept Study”. In opposing it to topical relevance, relevance is turned into a cognitive concept. A descriptor for this argument and knowledge claim could be “Cognitive Relevance Theory” revealing the implicit theoretical assumption of the theory and concept proposed. A methodological implication of this theory is its focus on the individual relevance judgments, which is why “Methodological Individualism” could be suggested as a descriptor too. That way the knowledge claim and its methodological consequences are indicated. Further, despite his critique of traditional retrieval testing, Harter emphasizes relevance as a measurable and empirical concept. This suggests an underlying epistemological assumption of Harter’s knowledge claim. An appropriate descriptor that would indicate this could be “Empiricist Epistemology”. As a consequence of psychological relevance, Harter also criticizes and rejects the validity of the results of recall and precision measures in applied in experimental evaluation and testing. Thus, a descriptor like “IR-Criticism” could be assigned since that would indicate the critique. This critique is part of Harter’s overall knowledge claim exactly because what the knowledge claim offered (psychological relevance) is the negation of what it criticizes. Since Harter aligns psychological relevance with various behavioral information seeking studies and argues that it underlies these, “Behavioral Studies” can also be suggested as a descriptor. This descriptor can be considered to be part of the argument because behavioral studies are a way of examining psychological relevance. Accordingly the descriptors suggested can be seen as a result of the epistemic-rhetorical situation Harter was faced with when writing his article. Exactly because they are part of this, we cannot expect that these be extracted on a mechanistic basis (e.g. by the use of automatic means). They form part of Harter’s knowledge producing activity and are as such embedded in a socio-rhetorical situation. The argument and the knowledge claims are something that must be contextualized in the social act of writing. Therefore, looking at the ways of invoking on, referring to, and evoking the contexts of the object under study, the literature of the field, the audience, and the author’s self can tell us what kind of argument is being made and in what the knowledge claims consist. These contexts reveal the very activity of the text of which it cannot do without. The activity of the text is situated in the activity of knowledge production. In terms of knowledge organization the above leads to the following claims: First, the reason why we can discuss the appropriateness of descriptors has to do with what arguments and knowledge claims a document put forward. This is again bounded to the rhetorical situation of the document and thereby its genre. Second, an understanding of a document demands more than its mere content in itself can explain; it demands an understanding of its activity as historically situated. The analysis presented above can provide a reasonable basis for discussing how the article is indexed in various scholarly bibliographies. A discussion of what the shift in genre implies. While the indexing of the article cannot be discussed in terms of
the truth or falsity of the indexing, the degree of appropriateness of the indexing undertaken can be discussed. Not all words and concepts are equally appropriate for an indexing of an article. In order to index scholarly documents and support the knowledge producing activity, knowledge organization must not fail the researcher (Weinberg, 1988). A scholar writing an article is faced with rhetorical problems in order to produce a knowledge claim. That way the analysis can be used to discuss strengths and weaknesses of knowledge organization and to what extent it contributes to cognitive authority, which may ultimately lead to specific descriptors stating a qualitative difference in relation to other access points. It may be objected that this kind of analysis is unrealistic because such a close reading cannot be done in 'real-life situations'. There may be a sense of truth in this. But then we will also have to admit that we cannot account for the shift in genre that happens when indexing scholarly documents in bibliographies and why that shift may pose a problem for searchers. We cannot expect then that scholars will use the systems for knowledge organization, as most studies show they do not. We cannot conceive of our knowledge organizing activities as 'user-friendly'. It is simply a chimera.

4. Practical Implications for Teaching Indexing

The problem of writing knowledge is discussed by Bazerman (1988). He poses the question of how students learn to write academic essays and what kind of problem writers are faced with: ‘Very soon into engaging this problem, I found that I could not understand what constituted an appropriate text in any discipline without considering the social and intellectual activity which the text was part of.’ (Bazerman, 1988, p. 4). As researchers and teachers of knowledge organization we face exactly the problem as Bazerman. The answer(s) to this question is exactly the same as those to the question ‘What do you need to know in order to index a document?’ It is of crucial importance for those having the task of indexing documents that they recognize the rhetorical situation they are in. This implies recognizing the social and intellectual activities documents are part of, or may serve, and that a choice must be taken as to what kind of descriptors are appropriate to assign in order to support the activity. This calls for a need of genre and activity knowledge. The indexer, like the informed writer (Bazerman, 1995), must know how to act within activities structured and shaped by the activities of the production, distribution and use of documents. These activities are historically constituted that have turned them into typified actions based in recurrent situations. This shows then a strong connection between how knowledge is socially organized through typified actions and how it is organized in information systems. The above analysis has, hopefully, revealed what we need to know about scholarly documents when organizing and representing them in information systems, and the limitations there exist when trying to ascribe cognitive authority to scholarly documents. Only then can the role of knowledge organization in scholarly communication be fully acknowledged and understood and the shift in genre from primary literature to bibliographic literature appropriately accounted for without a complete loss of cognitive authority. Only then does knowledge organization make a qualitative difference.
References
Augmenting Human Capabilities: Classification as Cognitive Scaffolding

Abstract: The argument presented here seeks to extend the notion of the classification scheme as a culturally-transmitted tool by emphasizing the cognitive value of the scheme's internal patterns of relationship. It elaborates on the use of classification as cognitive scaffolding (Jacob, 2001) and amplifies this idea through application of three constructs -- constraints, selections and expectations -- derived from Luhmann's (1995) theory of social systems.

1. Introduction.

The notion that mind is independent of the body is commonly understood as the hallmark of Cartesian philosophy. The mind/body dichotomy and Descartes's comparison of the human body to a clockwork mechanism has influenced machine functionalism and the functionalist claim that "The mind is a program run (in humans) with the brain as its supporting hardware" (Clark, 2001, p. 169; emphasis in original). There is increasing resistance to the Cartesian separation of mind and body and to the corresponding notion that cognitive functions consist, as Simon (1981) and others have argued, in the rational and mechanistic manipulation of symbolic representations. Nonetheless, Descartes's notion that machines and, by extension, living beings, could be understood through analysis of their constituent elements has had significant influence on modern scientific thought. As Capra (1996) observes, "The great shock of twentieth-century science has been that systems cannot be understood by analysis ... but only within the context of the larger whole" (p. 29). Thus, in contrast to the reductionist method of Cartesian analysis, a systems-based approach to structure would emphasize patterns of internal organization and the relation of the whole to the environment within which it occurs.

The argument presented here seeks to extend the notion of the classification scheme as a culturally-transmitted knowledge tool by emphasizing the cognitive value of the scheme's internal patterns of relationship. More specifically, it elaborates on the use of classification as cognitive scaffolding that was developed in Jacob (2001) and amplifies this idea through application of three constructs -- constraints, selections and expectations -- that are fundamental in Luhmann's (1995) theory of social systems.

2. Augmentation.

Engelbart (1963) presents a theory of augmentation that posits "the total system of a human plus his augmentation devices and techniques ... as a set of interacting elements rather than a number of isolated components" (pp. 1-2). He
argues that accomplishing the everyday tasks of living effectively in the world depends on external structures that extend basic human capabilities; and he posits an integrated system consisting of the Human using the Language, Artifacts and Methodologies in which he had been Trained [H-LAM/T]. The H-LAM/T system functionally combines the human agent and her culturally-transmitted tools in a synergistic unit such that any product of this integrated unit is greater than the sum of individual elements acting alone. Augmentation systems have emerged within a process of continuing sociocultural evolution that has generated explicit concepts, formal symbol systems, spoken language, written language, simple artifacts and complex technological machinery. Within this evolutionary process, the role of culture has been to provide the individual with a practical framework that integrates the human cognitive domain of language the concepts and symbols that comprise a culture's linguistic tools and the technological domain of physical artifacts so as to maximize the application of the individual's basic cognitive capabilities to the problems and goals of human society.

Clark (1997) echoes Engelbart's argument for the complex evolutionary advancement of augmentation devices when he describes the process as exhibiting a snowball effect: the enhancement of an existing tool, the introduction of a new artifact or the implementation of a new methodology allows the human to resolve a particular problem or achieve a specific goal, which frees him, in turn, to address other problems. But Engelbart (1963) contends that introduction of an artifact to address a problem situation may actually create new problems or point to new goals which lead, in turn, to the development of additional artifacts or methodologies. In this way, the process of technological evolution does not hinge on solving problems as much as it reflects the capacity of the individual, her cognitive tools and her culture to coevolve with the problems that confront the larger community. In this way, the ability of the integrated augmentation system to modify both its own orientation and that of the problem space is the ultimate achievement of the cognitive agent as embodied mind working in concert with his material and sociocultural environment.

3. External cognitive structures.

Within the augmentation system described by Engelbart, the cognitive agent and the physical and social worlds in which he participates evince a complex and interdependent relationship. Hutchins (1995) contends that the mind/body dichotomy in association with mechanistic attempts to model human cognition as abstract symbol manipulation actually generated an operating model of a sociocultural system from which the human agent was effectively excised. The real-world ability to use the language, artifacts and methodologies of the material and sociocultural environment is unique to the human agent actively embedded within his environment. By providing the tools and methodologies that support individual cognition and the production of effective, "rational" behaviors, the sociomaterial environment becomes an active extension of the human mind (Rumelhart et al., 1986).

Coupling of the individual with the material and cultural environment in which he participates is the essence of embodied cognition. Embodied cognition relies on complex external structures not only to distribute collective knowledge but to provide practical strategies for applying that knowledge. External cognitive
structures are physical, sociocultural or institutional tools that minimize the problem-solving efforts of the individual's day-to-day activities, either by prescribing institutionalized patterns of behavior or by providing the artifacts, technologies, knowledge structures or procedures that simplify interaction with the environment. As such, they serve to reduce cognitive demands on the individual by facilitating processes and operations, such as balancing a checkbook, that are not easily performed by internal cognitive processes alone.

Clark (1997) describes two external cognitive structures that function to facilitate and control human behaviors. *Stigmergic structures* (p. 186), such as traffic lights, institutionalize patterns of sociocultural behavior. These structures act as cognitive triggers to elicit patterned behavioral responses from the individual. While stigmergic structures inform and control individual cognitive activities, they are the residual products of successful problem-solving activities that have been undertaken by the larger sociocultural community. Like the system of traffic lights, they are often purposefully constructed in order to minimize the cognitive effort required to tailor individual behavior to specific events in the environment.

In contrast to the formulaic nature of stigmergic structures, *cognitive scaffoldings* encompass "a broad class of physical, cognitive, and social augmentations ... that allow us to achieve some goal that would otherwise be beyond us" (Clark, 1997, pp. 194-195). Clark contends that many cognitive activities rely upon or exploit the technologies, knowledge structures or methodologies developed by others which the individual then co-opts as cognitive scaffoldings to support his own problem-solving activities. This aspect of cognitive scaffolding is based, in part, on the developmental construct of the *zone of proximal development* proposed by the Russian psychologist Vygotsky. Vygotsky (1986/1934) theorized that there are certain points in the intellectual and social development of the individual when she is more receptive to the introduction of advanced concepts and procedures and can thus attain a significantly higher level of achievement by "leap-frogging" over intermediate stages of development.

Clark's notion of cognitive scaffolding extends Vygotsky's original idea to include not only the intellectual and social realms, but the physical realm as well. Culture provides the collections of tools that structure the material and social environment; the environment, in turn, gives shape to the human mind; and the human mind, in its turn, refines and extends culture (Clark, 2000). In this iterative process of interaction, refinement and extension, cognition is not the outgrowth of a solitary mind acting in a physical or social void but the synergistic product of a culturally-embedded individual actively and intimately coupled with his sociomaterial environment. Acting within the parameters established by the environment, the individual relies on a complex set of external physical, social, cultural and institutional structures that serve to reduce internal computational load, either by prescribing appropriate responses or by providing the scaffolding the strategies, procedures and tools that simplify problem-solving activities.

4. Constraints, selections and expectations in classificatory structures.

The scaffolding function of a classification scheme is impressive. It simultaneously orders knowledge and constitutes the basis upon which "reality" is constructed within a social system, knowledge domain or community of practice (Jacob & Albrechtsen, 1997). It supports praxis by determining the conceptual and theoretical content of the domain; by concretizing and maintaining conceptual
boundaries; by enforcing domain standards; by stabilizing meaning; by facilitating communication and the transmission of knowledge; and by continuing and extending the culture of the domain (Jacob, 1994; Jacob & Albrechtsen, 1999). The influence of the classification scheme is pervasive at the domain level; but its function as cognitive scaffolding is most notable at the level of the individual participant or practitioner. In some situations (e.g., Lakoff, pp.92-104), its utility as a cognitive tool may extend beyond the contribution of external scaffolding structure(s) to include the provision of standardized patterns of behavioral response which Clark (1997) identifies as stigmergic structures.

One way in which classification schemes serve as external cognitive structures is to place constraints on human understanding and activity constraints that reduce the complexity of cognitive tasks by simultaneously restricting the problem space and limiting the range of possible solutions from which the individual may choose. Whether they are posited in the form of universal principles (Chomsky, 1965), organizing structures (Gelman, 1990), guiding principles (Medin et al., 1990) or biases and predispositions (Markman, 1989), constraints limit the range of possible alternatives and provide selection procedures that guide the individual to favor the optimal alternative.

Keil (1981) argues that available evidence indicates that constraints cannot be reduced to a few basic organizational principles. He contends that forms of knowledge that involve less conscious reflection appear to be more strongly constrained: thus, for example, stigmergic structures such as the system of traffic signals are highly constrained, allowing little or no leeway for individual interpretation. As knowledge becomes more abstract and accordingly more complex, it may actually engender a greater number of cognitive constraints. Although the specific nature and function of constraints is undetermined, it seems obvious that, given the complexity and variety of cognitive activity, constraints do not simply impose limits on the range of cognitive possibilities but must also provide guiding principles for selecting the most probable alternative. Medin et al. (1990) point out that, "If organisms cannot, and therefore do not, examine all possibilities in some cognitive task, they must be 'prepared' or biased to learn some things rather than others, to draw some plausible inferences rather than others, and in general to favor some possibilities at the expense of others" (p. 169).

Constraints are generally formulated as restrictions on structure rather than process (Keil, 1990). Thus Nadel (1957) observes that "The parts composing any structure can vary widely in their concrete character without changing the identity of the structure" (p. 8; quoted in Luhmann, 1995, p. 283). That is, it is the pattern of relationships between elements, not the specific elements themselves, that constitutes the identity of the structure. At any point in time, each such relationship constitutes "a selection from a plurality of combinatory possibilities ... and only this selection can be held constant across change in elements, that is, can be reproduced with new elements" (Luhmann, 1995, p. 283; emphasis in original). It is the relationships between elements that serve as constraints and thus hold cognitive value for the individual. I know, for example, that Beamers will bark when the doorbell rings because he is a cairn and a cairn is a terrier which is a kind of dog and dogs bark at unexpected sounds; if I substitute Bleu for Beamers, I rely on the same selection of relationships and thus arrive at the same conclusion (Bleu will bark) because Bleu is a poodle; but if I substitute Bobo, I arrive at a different conclusion (Bobo will mew) because Bobo is a cat, thus prompting the selection of
a different set of relationships.

Without the constraints imposed by the relational structure of the classificatory system, the range of possible selections (e.g., bark, mew, moo, etc.) would be so broad as to make impossible any reproduction of the knowledge contained within the system. Only a structure that constrains relationships can acquire sufficient "internal guidance" (Luhmann, 1995, p. 283) to make possible the individual's production of knowledge about the environment: "From every element, specific other (not just any other) elements must be accessible, and this must be so due to specific qualities of the elements that stem from their own accessibility. To this extent structure as the selection of constrained possibilities is presupposed in the constitution of qualified elements ... the constraint [is] on the quality and connectability of elements" (Luhmann, 1995, p. 283).

Constraints facilitate the selection of certain knowledge-bearing relations over others and in this manner function as expectations. Because expectations are generated by the intermediate selection of a limited set of possibilities, they condense knowledge at two levels. First, they condense the referential structure of the classification scheme and ease "the burden of selection" (Luhmann, 1995, p. 96), allowing for quicker selections of appropriate responses. Second, they condense symbolic generalizations, which "stamp identities onto the flux of experience" (Luhmann, 1995, p. 94) by defining objects, events and concepts. Symbolic generalizations are "contained and refabricated within a network of expectations. They organize or better, continually reorganize expectations, and, depending on the course of experience and action, they take up material from the underlying referential strata of meaning complexes or allow what is too seldom used to sink back down" (Luhmann, 1995, p. 96). Working simultaneously on these two complimentary levels, expectation "at once constrains what is possible and makes visible other possibilities ... [and] leads to the emergence of structured complexity (organized complexity)" (Luhmann, 1995, p. 97; emphasis in original).

Classification structures are sociocultural structures that organize the complexity of the environment. They are, as Luhmann, observes, "expectational structures" (1995, p. 292) that constrain the range of possible relations by linking elements within the structure. Indeed, Luhmann (1995, p. 292) claims that expectations are the constraints that determine not only permissible relationships within the classificatory structure but also the organization of the structure as a whole. As Buckley (1967) points out, "The structure of such a system is then viewed in terms of sets of alternative actions or tendencies to act in certain ways, associated with the components and the constraints that specify or limit these alternative actions. The genesis of organization is thus the generation of these sets of alternatives and the constraints defining them" (p. 128; quoted in Luhmann, 1995, p. 566; emphasis in original).

5. Conclusion.

Explosive growth in the number and variety of digital resources has underscored problems of comprehension and retrieval associated with the lack of physical structure on the Web. Faced with the need to construct information architectures that will promote user understanding, facilitate navigation and provide efficient access, information professionals have rediscovered principles of representation and organization that traditionally provided a foundation for systematic organization. But the development of effective information systems
hinges on a comprehensive understanding of social and cognitive environments as well as knowledge of organizational structures and technological capabilities.

The notion of classification as cognitive scaffolding has the potential to contribute significantly to an understanding of how individuals and communities use knowledge structures to augment their basic abilities. A classification scheme functions as an external cognitive structure on several different levels. At its most basic, the scheme provides a simple stigmergic structure that reduces cognitive processing by prescribing appropriate responses according to institutionalized standards. At a higher level, it operates as a constraint-satisfaction network (Clark, 1997) to prescribe class membership. At a more complex level, the scheme orders and thereby simplifies the individual's experience of the environment by condensing knowledge into sets of expectations and by constraining the knowledge-bearing relationships that obtain between classes. Application of Luhmann's (1995) theory of social systems shows promise for extending current understandings of how classification functions as cognitive scaffolding on both the individual and social levels.

References.


Universal Concepts, Cultural Warrant, and Cultural Hospitality

Abstract: The problem of how to provide access to information regardless of linguistic or other domain boundaries or cultural traditions can be approached by examining how cultural universals are implemented in specific cultures at specific times and places. The universal concept of “time” and its implementation in calendars is used as an illustration, and how time is treated in knowledge organization systems is briefly described. A broadened definition for the concept of “hospitality” is proposed for use in evaluating the efficacy of knowledge organization systems. The identification of the complementary concept of “cultural hospitality” provides a theoretical framework to inform decisions about the types of access that can (and/or should) be provided by knowledge organization systems that purport to be globally useful and ethically balanced.

1. Cultural Warrant and Conceptual Universals

The concept of cultural warrant posits that every classification system is based on the assumptions and preoccupations of a certain culture, whether the culture is that of a country, or of some smaller or larger social unit (e.g., ethnic group, academic discipline, arts domain, political party, religion, and/or language) (Beghtol 2001a, 1986). The concept of cultural warrant implies that a knowledge organization system is more likely to be useful and appropriate for those who are members of a culture and that it is less likely to be useful and appropriate for those who belong to some different culture, at whatever level of society that culture or domain may reside. Thus, a knowledge organization system that is appropriate for the elements of one culture may not recognize elements that are highly important for some other culture, and such exclusions pose problems because we need to integrate knowledge across cultural, geographic, and linguistic boundaries. Important tensions thus exist between culturally specific knowledge organization systems and the need to provide systems that are both globally accessible and culturally appropriate. In general, knowledge organization systems for global usage would need to incorporate all the various syntactic and semantic foundations of any and all of the world’s cultures, and the creators of knowledge organization systems need to create techniques for poly-cultural information retrieval. Since the foundations of different cultures may be in conflict with each other, we need to develop theories and techniques for incorporating any possible cultural assumptions that might be used for information retrieval.

To gauge the extent of this problem, it is useful to examine relationships between universal semantic and lexical concepts and the specific implementation of those universals in different cultures. Goddard and Wierzbicka (1994) identified eight sets of semantic and lexical universals: Mental predicates; Determiners and qualifiers; Actions and events; Meta-predicates; Time and place; Taxonomy and partonomy; Evaluators and descriptors. Some of these have been examined for their
various roles in bibliographic classification systems (e.g., Breghtol 2001b, 1997). In this paper, the universal concept of time and its instantiation in calendars is used to illustrate the complexities of providing culturally specific information for global information retrieval requirements and to analyze the problems of providing culture specific information resources in globally and ethically acceptable information retrieval systems. Calendars are an appropriate example because they subdivide the abstract domain of "time" for cultural purposes and because detecting bias in abstract domains is more difficult than detecting bias in other domains (Brey, 1999).

2. Calendars

Calendars are non-universal cultural artifacts that represent and help to create the various syntactic (structural) and the semantic (meaning) aspects of a society. Among other things, a calendar structures a person's daily and legal life and the whole culture's yearly, monthly and seasonal cycles. It sets the times for secular festivals, religious occasions, and civic holidays. Calendars are thus a knowledge organization system for various kinds of cultural knowledge. For this reason, the use of a particular calendar is a significant choice for a culture, and calendar reform is regarded as a serious intrusion into established cultural traditions because a new calendar changes a culture. For example, a "calendar war" (Edmonson, 1976, p. 716) can erupt when one culture conquers another, and advocates of the astronomical accuracy of a one calendar over another have been executed for their beliefs (Chu, 1997). Modern movements to reform the calendar have failed for various reasons (e.g., Davies, Trivizas & Wolfe, 1999), and efforts to establish a uniform calendar for the whole world do not seem to have created much interest (e.g., Volk, 1995).

Despite the cultural specificity of different calendars, the globalization of electronic information has revealed that the ability to "translate" one calendar into another is important in a number of fields (e.g., business, banking, computer science, historical research). Dershowitz and Reingold (1997) created computer algorithms for comparing fourteen calendars to each other: Gregorian (civil calendar); ISO (commercial); Julian (old civil); Coptic; Ethiopian; Islamic (Muslim); Persian (modern); Bahá'í; Hebrew (Jewish); Mayan (2); French Revolutionary; Chinese; Hindu (Indian, old, mean); Hindu (Indian, new, true). Some of the difficulties of this enterprise may be appreciated by referring to the Anglo-American Cataloging Rules (AACR2), which describes how to convert the Julian to the Gregorian calendars at different times for different countries for the purposes of bibliographic description (AACR2 1998: 415, fn 17).

3. Time and Calendars in Knowledge Organization Systems

The provision of different calendars has not been a priority for classification and knowledge organization systems, presumably because the cultural warrants of these systems have not required access to different calendars, although chronological arrangement of books on shelves has been attempted at least since the invention of Biscoe Time Numbers for science in 1885 (Lehnus 1980). The only suggestion that users might want different calendars appears to be in the ontology-
The most complete treatment of time appears in the *Universal Decimal Classification* Table Ig Common Auxiliaries of Time. In Table Ig, "the basis of date indication is the Christian calendar, but non-Christian systems of time reckoning are also allowed for...as well as other time concepts, e.g., seasons and geologic time" (1993, Part 1 p. 66). Table Ig is undergoing revision, but the possibility of converting all dating practices to non-Gregorian calendars is apparently not under consideration (Robinson 2000). Similarly, other knowledge organizations systems such as thesauri, (e.g., *Art and Architecture Thesaurus*), subject heading lists (e.g., *Library of Congress Subject Headings*), and other kinds of metadata systems such as ontologies generally assume the use of the Gregorian calendar and have not given non-Gregorian calendar options.

This kind of examination of the provisions made for one linguistic and cultural universal in knowledge organization systems needs to be repeated for the other linguistic and cultural universals identified by Goddard and Wierzbicka (1994). As an example, however, this study of provisions for calendars shows that current systems are unlikely to be able to accommodate the information needs of users who belong to different cultures in their preferred calendrical mode.

4. The Concept of "Cultural Hospitality"

In order to develop knowledge organization systems for globalized information access and retrieval, we need a theoretical framework for knowledge organization systems that will privilege the needs of different cultures, whether they are national, ethnic, domain or disciplinary cultures. The concept of "hospitality" has long been established as one of the desiderata for the notation of bibliographic classification systems. Hospitality is the ability of a notation to admit new concepts appropriately and to accommodate them in the correct relationships with other concepts. To develop a new theoretical framework for knowledge organization systems, it is helpful to expand the concept of hospitality by 1) including provisions for hospitality beyond notational issues and 2) broadening hospitality to include different cultures as well as new concepts. These extensions of hospitality provide us with a new concept, "cultural hospitality".

Cultural hospitality complements and extends the idea of cultural warrant. It posits that making provisions for specific aspects of different cultures in knowledge organization systems will increase the appropriateness and usefulness of those knowledge organization systems in different settings for the purposes of world-wide information flow. The fundamental tenet of cultural hospitality is that knowledge organization systems should be "permeable" (Olson 1996, p. 9) to different points of view and different cultural attitudes and practices. Practical implementation of this ideal may not be easy. Like all theoretical frameworks, however, the concept of cultural hospitality provides a goal toward which we may proceed, and some possible methods of attempting to reach this goal are discussed elsewhere (Beghtol 2002). Thus, the ideal of cultural hospitality provides a method for assessing the acceptability of provisions made in knowledge organization systems for different cultures, at whatever level of society these cultures may reside.
5. Ethical Knowledge Organization Systems

Identification of the concept of cultural hospitality allows us to begin an investigation into the ethical design of knowledge organization systems for worldwide applicability. Brey (2000) set out a process for developing an ethical basis for computer systems that included three levels:

- the disclosure level, at which analysis of systems on some ethical value takes place;
- the theoretical level, at which core theories are identified and refined; and
- the application level, at which the theories from the second level are applied to disclosures made at the first level and potential implementations are tested and put into practice.

This model of the process of ethical development for computer systems applies equally to knowledge organization systems. Some research has been undertaken on each of these levels, including efforts to disclose embedded values in knowledge organization systems, to internationalize systems, to create poly-lingual systems, and to design and implement applications that can underpin and promote these different efforts. One possible high level ethical system that could be used to unify these activities would be the United Nations Universal Declaration of Human Rights (UDHR, accessible at http://www.unhchr.ch/udhr/lang/eng.htm) combined with Smith’s concept of Global Information Justice (GIJ) (Smith 2001). These are appropriate foundations for the ethical assessment of knowledge organization systems because they have been developed on the assumption that differences among cultures should not be threatened by the globalization of information and information technology and that preservation of cultural differences is a high priority (Beghtol 2002).

6. Conclusion

From an examination of the concept of cultural warrant and of how one universal concept plays out in a cultural artifact such as a calendar, we have derived the theoretical principle of cultural hospitality. This concept needs to be debated, assessed and tested further to assess its potential for effective implementation. It appears, however, that we need to make specific universal concepts available for appropriate uses in different cultures and for different information tasks, and that this endeavour can promote the development of ethically-based knowledge organization systems.

Acknowledgements

This research was partially supported by the Social Sciences and Humanities Research Council of Canada grant 410-2001-0108. Research assistance was provided by Anna Slawek and Ann Simonds.

References


For centuries, there have been few doubts about the validity and the nature of the principles that put together knowledge, texts, documents, disciplinary subject matters and the respective literature. Kochen, in the sixties, declared that literature organizes knowledge as knowledge organizes the state of affairs in the world. And, regardless of how information classification and indexation will be defined, they did not intend to interfere, through their technical-intellectual operations and tools, into the canons that rule specific knowledge. On the contrary, the more perfect the reversibility and neutrality of the documentary operation, the greater the effectiveness of its intervention.

In the late nineteenth century, however, the premises that assured the universality of knowledge organization principles were not so trustworthy as they used to be. Similarly, later on, the conventional and selective character of documentary operations was acknowledged. But only in more recent decades, previous forms of linkage among subjects, knowledge, discourses, and documents have experienced decisive changes.

These transformations become more noticeable and intensive in the electronic communication networks: a) permitting the reversibility of roles, such as receiver-issuer, writer-reader, producer-consumer; b) rendering less stable and distinct the institutional patterns, actor identities and previous shared ontological premises.

It is possible, anyway, that information links replaced, in the mediating networks, the social bonds of community and social actors embedded in a relatively autonomous and unificated field: a specialization or discipline, an economic activity, a local community?

We can ask ourselves, in these contexts, about the conditions of "informational seeking" at the new environments, when cross-retrieval and seeking situations are incremented.

Selective process and decision making about the informativity of a source, text or document, presumed a judgement of values, whose parameters and criterions of selection and adherence must change in different contexts of action.

Information domains are embedded in the life world of sense production. This "world of life", an original space of sense making, is the meeting place where are submitted, as decision making instance, dissidents and alternatives practices and discourses. In these conditions of arbitration space, a "life world" is a nearly transcendental level, where the speakers and the hearer put in common claims of validations. This arbitration sphere is not only a source of legitimization, but also a framework of building collectives "programs of actions".

Geertz (1988) names "world life" all communities that shared, during long
periods, experiences and common situations: professions, scientific communities, activities partners.

With these premises, the world of life is the framework of the selective information seeking, and the decision-making about the heuristic, argumentative or demonstrative value of information source. Information, in this context, has a testimonial status, in social process that claims for arbitrate resolutions and demands justifications’ arguments: political, juridical, inquiring practices.

Now, however, there are indicatives of other trends, a breakdown in the “communitary” model. The information seekers, in network context, are groups of complex relations that search for external resources of information, at the same time similar and different in any aspect, for any participant, in certain opportunities, according to a certain variable.

At the same time, it would seem that when a person is included in an electronic network of information access, where a huge volume of data and documentary inscriptions is being offered, he has no equally easy access to the information about the information that could provide him a better practical and epistemological use of these resources.

In this framework, we are asking about the selective behavior of individual and collective subjects in communication networks, and their actual conditions of choice information sources, subject matters and forms of knowledge.

Do new technological and cultural alternatives result in more possibilities for selective choice of information sources, services and products? Or do the actual deregulation of intellectual and discursive frames have effects on anomie immediately?

In summary, a main question is about social subject’s autonomies or anomie behaviors in cross-retrieval situations- between transversal and heterologic relations, with different institutions, thematics and epistemological fields.

“Durkheim defined the term anomie as a condition where social and/or moral norms are confused, unclear, or simply not present. Durkheim felt that this lack of norms—or pre-accepted limits on behavior in a society--led to deviant behavior.” (The Emile Durkheim Archive, 2002)

In what aspect does the information network assume the implicit requirements of autonomous subjects and universal access to public knowledge and information sources? Or, on the other hand, do the new technologies of communication and information increment a new illiteracy and information anomie?

The transformation of information medium and background are now affecting the previous systems of inscriptions of meanings and the marks and attributes of "indexicality", previously defined for documentary inscriptions. The rescue of indexicality by meta-data encapsulated document will be not a sufficient support for the information user autonomy. The digital environment, at this moment, has a tendency to lose context components and cross-references. But these are important factors for the text’s interpretation and for access and apprehensions of informative semantic contents, in a given domain.

According to this, it is possible to speak of information anomie tendency. Then, when more abundant and lessy the information flux is, the more significant the desinformation effects may be for numerous legions of less favored people and for numerous countries with less control of the information meta-environment. In a era of hegemonic digital voice, is augmented a new digital illiteracy augments --
because many literacy groups are losing their previous competencies and haven’t access to digital writing. Information anomie, also, will be the effect of a non-activated communicational contract, in the cyber space, because of the actual absence of numerous actors in the national and international forums where the dominant rules and strategies of information regime are stipulated.

Building a concept of the autonomous information subject, it is possible to think of the scientist as a gnoseologic actor, exemplary of relative autonomy and self-government in the generation and use of information.

The research activity is rich in tacit and explicit selective processes, in seeking internal and external information sources. The inquiring act defines its self-domain of intelligibility. In this sense, the evidence of information sources are constituted in the course of asking; when that was not a problem, it turns over a problematic matter and, what haven’t previous links, now obtain important relations. In the most important part, this control of the knowledge process is generalized in handbooks and in methodological rules.

The scientist is a competent worker of knowledge and meta-information in heterologic networks, and has a significative informational mobility, as a result of their conviviality in rich information and meta-information environments.

What characterizes autonomous inquiring, in this point of view, is not their insulation among peers, but the competent use of meta-cognition and meta-knowledge mechanism, the availability of literary technologies and the competence for its use in the credibility cycle building. But this requests the maintenance of meta-environment of controlling and monitoring knowledge productions, based on indexical proprieties of cultural artifacts and symbolic mediations. These are contributions to knowledge translations in cross-disciplinary and multi-activity fields. Science research programs are not a result of spontaneous exercise of human faculties, but a cultural practice that needs an explicit and collective project of development.

In this way, meta-information (or second level of information) foresees and precises the possibilities of information usage for knowledge production and, according to this, is one of the necessary conditions for cross-boundary knowledge integration.

To access the scope and relevance of these assumptions, this paper intends to reconstruct: a) the ways collective subjects build their classifying ontologies and its forms of knowledge and information validation, according to their "pragmatic differentials" and "evidence cultures" – in specific and real action contexts. b) the procedures that establish new links and relations for interaction in a heterogeneous communication networks; c) the answers and criteria with which the interfaces with long distance interactive and computerized communication network, are operated.

References


Subject Ontogeny: Subject Access Through Time and the Dimensionality of Classification

Abstract: Classification schemes undergo revision. However, in a networked environment revisions can be used to add dimensionality to classification. This dimensionality can be used to help explain conceptual warrant, explain the shift from disciplinary to multidisciplinary knowledge production, and as a component method of domain analysis. Further, subject ontogeny might be used in cooperative networked projects like digital preservation, online access tools, and interoperability frameworks.

1 Introduction

Knowledge changes through time. Classification schemes as tools for accessing knowledge undergo constant revision. It is impossible to claim that the ontology of subjects and their interrelationships, once established by a classificationist, remain constant within that scheme. As revisions to classification schemes emerge, so too do new subjects. These, new parts of the updated classification scheme are elements in a formal system – elements that represent the current interpretation of knowledge. The classes in a classification scheme are mutually exclusive and jointly exhaustive. Each class can be seen as a description of what the other is not, within this system. Thus, as the classificationist's interpretation of literary warrant changes, so too do the classes, their boundaries, and their interrelationships within the scheme. Normally, the result is a revised scheme absent of any record of what the knowledge landscape looked like before this current scheme. In a print medium, the record is set in type, allowing the curious to research the interrelationships and terminology of knowledge as it has grown. In a digital networked environment, a scheme can be constantly updated. And with each revision the palimpsest of the past paradigmatic cosmos disappears.

The following example illustrates how the DDC's interpretation of knowledge has changed over the last century. Eugenics, (a term with various definitions through the Twentieth Century) has been classed in the Dewey Decimal System alternately at 575.1, (with Genetics in before the 16th edition) and 363.92 (under Social Problems and Population). DDC 20 provides an index entry for Eugenics that looks like this:

<table>
<thead>
<tr>
<th>Eugenics</th>
<th>363.92 [Population quality]</th>
</tr>
</thead>
<tbody>
<tr>
<td>crime prevention</td>
<td>364.4</td>
</tr>
<tr>
<td>see also Crime prevention</td>
<td></td>
</tr>
<tr>
<td>health</td>
<td>613.94</td>
</tr>
<tr>
<td>medical ethics</td>
<td>174.25</td>
</tr>
<tr>
<td>see also Medical ethics</td>
<td></td>
</tr>
<tr>
<td>population control</td>
<td>363.98</td>
</tr>
<tr>
<td>social services</td>
<td>363.92</td>
</tr>
</tbody>
</table>
sterilization services 363.97

The entry in DDC 20 for Genetics is:

Genetics 575.1
animal husbandry 636.082 1
animals 591.15
humankind 576.139
social theology 291.178 365
Christianity 261.836 5
sociology 304.5
microorganisms 576.139
plants 581.15

The Relative index of the 16th edition reads as such:
Eugenics 613.94 [same as Eugenic practices—hygiene]
 formerly *301.323; † 575.1

The entry for Genetics in DDC 16:
Genetics
animal 591.15
general 575.1
human 613.9
plant 581.15

Above is an example of a subject ontogeny.

Current interest in the human Genome project will arouse curiosity in readers. They may ask how documents classed in 575.1, placed on the shelf, marked either as Eugenics or Genetics by classed on this topic relate to what is being discussed in the news media, for example. It may also be true that Eugenics is still part of Genetics, as a discipline, in some relationship other than represented in the class numbers above. What kind of access is granted by a classification system that shows how knowledge has changed, verses one that revises classes, denying access to the classificationist's interpretation of the change in knowledge? With each revision, a scheme for classification cuts itself off from its previous view of knowledge, building an artificial boundary of time. There are other rhetorical questions pertinent to time as it relates to subject access. For example, could one access the array of subjects in higher education that were taught during Plato’s Greece? Through a classification scheme, can one collocate the works of proto-anthropologists? These knowledges are not reflected in classification schemes, because each living scheme needs to be revised to be viable – thereby eliminating the fossil record of literary warrant. To what degree do revised classification schemes blind us to how subjects change and are re-collocated through time? What can knowledge organization theory do to help the sophisticated user re-collocate knowledge through time? This can be answered by charting the development of a class in a classification system through time. In other words, this can be answered by charting the subject’s ontogeny.
2 Classification Theory and the Dimensionality of Classification

Discussions surface constantly about the poverty of classification systems. Not necessarily the schemes themselves, but the conceptual accompanying material, the types of description they are not successful at, and on what philosophical ground they are based. Each of these complaints, outlined in further detail below, points to the social and documentary nature of classification schemes. That is, classification is a social and documented practice. Thus, as people interact (in a social way) with this information structure they are using, in a large way a document – the classification scheme. There is often a disconnect however, between the social dimension and the documentary dimension of classification. Cochrane, Beghtol, and Hjørland outline three examples of this disconnect.

2.1 Conceptual Warrant

In an effort to aid online searching Pauline Cochrane, asks for a tool, “some kind of management information system which would collect data about concepts indexed in our databases and provide some structured analysis” (Cochrane, 1995 p. 36) that will allow us to manage the semantics of the system. That is, she wanted to see a tool built for information professionals to manage the classification system as it grew and changed to reflect literary warrant. This tool would be used to fix the extension and intension (breadth and depth of coverage) of class numbers as warrant demanded. This same tool might prove helpful for us to examine how literary warrant required is shaped through time, providing access to older conceptions of the universe of published materials.

2.2 Organizing Principles of Classification Systems

Clare Beghtol (1998) questions the organizing principles of knowledge organization schemes. They are: disciplinary structure, fiction/non-fiction distinction, and the document as a unit of analysis. The functionality of current classification schemes is questioned here because 1) they are not flexible enough to express meaningful relationships between, within and among disciplines, 2) they divide the universe into narrative and non-narrative dichotomies though literary warrant does not reflect this division, and 3) current schemes do not express the “internal elements of documents,” (Beghtol, 1998 p. 2). However, each of these principles fit with the common interpretation of the universe of knowledge – at one time. They no longer fit the interpretation of knowledge organization because of current literary warrant and information technology. Were mechanisms in place for subject ontogeny to adapt to these shifts, without rupturing other economic and social mores, for example in database displays, the expressivity and flexibility sought by Beghtol might be achieved. In this sense subject ontogeny is much like the need for multiple views outlined in Beghtol (1998) and Albrechtsen and Jacob (1997).

2.3 Domain Analysis

Related to the idea of subject ontogeny, Birger Hjørland (1998) establishes a rationale for classification to examine knowledge as it changes through time. This is the historical method, one of four basic kinds of methods (Hjørland, 2002). The historical method is one way of understanding the story of a domain. It is one way of understanding that the “classification of a subject field requires a concept or view of that particular field,” that a “classification cannot be neutral regarding its approaches or theories about subject matter,” (Hjørland, 1998 p. 164). This, like a shift from disciplinarity to multidisciplinarity as a design construct, can be documented in mechanisms that manage subject ontogeny. Thus, as a part of the
tool of classification, subject ontogeny might be used as tool to refine classification via methods of domain analysis.

3 Mechanisms for Subject Ontogeny

The mechanisms for a subject ontogeny are at our disposal. Some are in place, though not used in a robust way. Authority files are one example of the way that revisions are managed now with subject headings. That technology is present. What is not present is the expertise and desire to shape this historical perspective in classification. What is needed is a simple tool that can be added to displays, that offers a representation of the subject ontogeny.

4 Applications

4.1 Metadata Preservation Models

The OAIS model (CCSDS, 1999) is a model, widely adopted in the digital preservation community. This model wraps digital data with metadata for future access. That future access, even if based on an established system, may hinge on the consciousness of subject ontogeny to the digital objects contained in this model. That is, in order to retrieve, through time, this kind of digital object, systems can be built that allow for updating and linking classes to the predecessors. By linking classes, as opposed to revising completely (that is deleting the old class), the information system can preserve relationships as they existed before.

4.2 Online Access Tools

Marcia Bates argues for a distinction between indexing and access for networked information resources, (Bates, 1998). The distinction she draws between indexing and access puts knowledge organization in the middle of the two. She says, "the user front-end can be designed around the distinctive traits and evolutionary adaptations of human information processing, while the internal indexing describing the document may be different. Changes in our developing understanding of human cognition in information-seeking situations, and changes in vocabulary can relatively quickly be accommodated in a user-oriented front-end, without requiring the re-indexing of giant databases," (Bates, 1998 p. 1202-1203). That is, if the database is designed with the idea of subject ontogeny in place.

4.3 Interoperability

Current work in Digital Libraries is looking ardently at how collections, and specifically their metadata can be shared across the network — how they can interoperate. Many manual and automatic solutions are being suggested (Arms et al., 2002). What is clear from this work is coordination in the online environment requires policy that shapes knowledge organization structures’ ability to evolve. Subject ontogeny, as a matter of design will aid in work load in such projects.

5 Summary of Classification’s Potential Dimensionality

If we take a class to be an object in a system of objects that have relationships (hierarchical or otherwise), we can, by a number of warrants, chart its development through time. Subject ontogeny augments the dimensionality of classification schemes.

That classification is an interpretive process is not in question. Placing items
in relationship to one another is an act of interpretation. However, the dimensions of those relationships need to manifest in information displays. Cochrane, Beghtol, and Hjorland call for an expression of dimensions. A user curious about the history of Genetics, clicking through a display using DDC numbers may seek an explanatory display about the history of this class, 575.1.

5.1 Encyclopedism

That classification needs dimensionality expressed in its construction, display and initial analysis, is evidenced in the literature cited above. By acknowledging this need, and by positing corrections and additions to this *document* that is a classification scheme, we are moving toward an encyclopedic structure. An encyclopedic structure is one that tells us many things about the nature of subjects. It tells us about their origins, how they change through time, and if they have ceased to exist. Again, this technology and interpretive stance is present in knowledge organization writ large. However, there has yet to be an expression of these needs in a display that allows us to access the social life of the class through a classification scheme or a class number.

Subject ontogeny, if added to classification, will address concerns classification theorists have expressed in the literature for a change in classificatory structure. Subject ontogeny could be used to express ideas of conceptual warrant and how those change through time (Cochrane, 1995). It can help us interpret multidisciplinarity, while recognizing the established institutions of knowledge production (Beghtol, 1998). Further, subject ontogeny will allow us to express the historicity of subjects in domains, (Hjorland, 1998). Most importantly, subject ontogeny is not a radical change, but an addition to existing systems. It is an interpretive layer that can serve the needs of information workers as well as users.

Notes

1 The social commentary offered by the DDC is well documented elsewhere. Any discussion of this context is beyond the scope of this paper.

2 However, because classification can be seen as a document (in the abstract – as a construct of an intellectual pursuit and in the concrete – the form of written schedules), it is therefore an artifact. Classification is a fossil in the sense that it is a record, no longer is it an access tool. Classification, in all its successive versions, could be considered both a record and an access tool, if we take into account the idea of subject ontogeny.

References

Albrechtsen, H. and Jacob, B. (1997). Classification Systems as Boundary Objects in Diverse Information Ecologies. In E. Efthimiadis (Ed.). *Advances in Classification Research*


Merging Ontologies using a Bottom-up Lexical and Structural Approach

Abstract: Information integration has been the focus of different research domains for several years. With the emergence of the Internet as a very large database, this topic has become of extreme relevance for a larger audience, seeking mechanisms to align information from different data sources according to semantic needs and constraints. The classical integration approaches do not satisfy new operational requirements; thus new strategies should be proposed and developed. We suggest the adoption of a light alignment mechanism without merging data source. The alignment process occurs at an ontological level, setting the components and transformation functions necessary to translate data from source to target. Establishing the alignment process at the ontological level allows the system to reason about the semantics embodied in the information and thus enhance the alignment results, according to both the source and target specifications. The resulting specifications are then applied in the transformation procedure, executed independently of data source entities.

1. Introduction

The Semantic Web vision proposes the representation of data semantics and their exploitation in a machine processable way, contrasting with the current user-based approach, characterized by the discrepancy between the huge amount of existent data and the user capabilities to transform it in useful knowledge. Many standards (Semantic Web, 2002; W3C, 2002; UDDI, 2002) have been proposed for the semantic web and knowledge interoperability, but entities have different information conceptualizations, consubstantiated in different information systems and respective ontologies (Guarino, 1994; Ushold and Grunninger, 1996) that can not be solve by the adoption of these standards Accordingly, information integration research (Pinto, Gómez-Pérez and Martins, 1999; Bergamaschi et al., 2001, Madhavan, Bernstein, and Rahm, 2001; Critchlow, Ganesh and Musick, 1998; Neumann et al., 2001; Visser and Tamma, 1999; Doan et al., 2002; Park, Gennari and Musen, 1997; Rahm and Bernstein, 2001), experienced new motivations and a very challenging application scenario. In addition to the classical alignment problems, requirements in the Semantic Web are augmented due to decentralization (many entities describing eventually the same thing), dynamics (providers, consumers, content are always changing) and decontrol (no entity supervises the Web content). Typical alignment processes are very time consuming and the focus is mainly on accuracy and operational performance. We suggest that alignment process in semantic web should be fast and computationally inexpensive, and must focus on the relevant information components for each specific interoperability (Silva and Rocha, 2002). Additionally, we believe that in many situations the accuracy dimension of the alignment is not so fundamental and can be relaxed.
In the next section we propose a conceptual alignment process describing the envisaged phases and propose different approaches for their operation. In section 3 we describe our understanding about the semantic bridge concept, both upon the process and the resulting specification. A description of a small case allows us to present examples of our proposals. Section 4 summaries the current work and the main contributions, and suggests future steps.

2. Alignment Process

According to previous considerations we now introduce and describe the identified phases in the ontology alignment process.

The normalization phase respects the processes whereby both ontologies are synchronized to unifying language and structural model. Additionally, minor lexical normalizations are executed. The former concerns the constructs and may cause some changes in the represented knowledge, while the latter deals with content (e.g. expansion of acronyms and abbreviations) and should avoid changes that might imply semantic modifications. In both cases it permits maximizing process abstraction (Omelayenko and Fensel, 2002).

The similarity measuring phase is the set of processes that specifies similarities between entities from both ontologies. Four different approaches exist in measuring ontology similarities:

- Syntactic analysis is concerned with measuring similarities between elements names. Simplest approaches consider names as set of characters (Levenshtein (Maedche and Staab, 2001) edit distance) but more complex ones attempt to infer syntactic relations between words. For example, consider the words “may” (month and simple present of verb “to may”) and “might” (“power” and simple past of verb “to may”). The Levenshtein edit distance would result in a similarity of 4 (poorly similar), while a syntactical analysis would mention the relation centered in the verb sense. Although, it is difficult to translate human readable information into computer quantifications;

- Lexical analysis measure similarity according to lexical relations and semantic meanings. Typical approaches use thesaurus, glossaries and other lexical tools like WordNet;

- Structural analysis consider the organization of (and relations between) elements in the ontology. Normally, an ontology has an underlying taxonomy, which is a good source of information for similarity measuring. Clustering techniques, based in syntactic and lexical similarities are typically used in this analysis (Bergamaschi et al., 2001);

- Extensional analysis evaluates similarity between instances of different ontology elements, which permits evaluating a similarity of elements according to their instances.

The bridging phase intends to define which entities from the source ontology are semantically transposable with which entities in the target ontology, and how the instances associated to the first are transposed to instances in the second. Thus we distinguish two substantially different sub-phases, although closely related:

- The association sub-phase takes as input the similarity measures
calculated in the similarity measuring phase, and defines associations between elements from both ontologies;

- The rules specification sub-phase is concerned with specifying, for each association, the process whereby the instances from the source ontology can be transformed into instances of the target ontology. In the simplest case the necessary function would be just a copy of values, but in other cases complex rules should be applied in manipulating several entities from both ontologies.

The representation phase intends to represent semantic bridges using an external representation language. This representation is necessary when the execution phase is performed by an entity other than that which specified the bridges. Hence, this phase is not mandatory if the bridging and execution phases are carried out by the same entity. Nevertheless, in order to reuse those bridges it is recommended to represent them in an external, eventually widely accepted standard formalism. However, such language, besides a minimal proposal (Omelyanenko, 2002), does not currently exist. A brief description of its requirements is presented in Section 4.

The transformation phase intends to effectively translate instances from one data source to the other, according to the semantic bridges specified in the bridging phase. There are some relevant initiatives in the transformation field in the context of the semantic web, namely XSLT (W3C, 2002) and TRIPLE (Sintek and Decker, 2001). The former is syntactic oriented, while the latter is an independent representation interpreter and rule processor. Its reasoning capabilities can be further expanded by connecting it to external inference and reasoning systems.

The transformation phase should be responsible for guaranteeing some quality in the resulting instances. This quality is hard to measure but some of its dimensions are: (i) unique instance identification (prevent identity duplication and same instance with different identity) and (ii) correct categorization of instances. Figure 1 presents a simple example of categorization refinement: at ontology level it would not be possible to define semantic bridges between instance values, thus the value of Course attribute would not be filled if extensional analysis could not be performed. Later we describe a specific formalism that permits expanding the ontology.

The negotiation phase deals with the process whereby the "owners" of entities try to achieve a consensus about the semantic bridges. The negotiation process depends extensively on two components: the negotiation protocol and the arguing capabilities of the entities. Minimal research exists in this domain and none relating argumentation. The negotiation process may assume two distinct models according to the moment the negotiation occurs: (i) Progressively (for each derived

---

**Figure 1** - Ontology view Vs. Instance view: data may complement semantic bridges

---

The negotiation phase deals with the process whereby the "owners" of entities try to achieve a consensus about the semantic bridges. The negotiation process depends extensively on two components: the negotiation protocol and the arguing capabilities of the entities. Minimal research exists in this domain and none relating argumentation. The negotiation process may assume two distinct models according to the moment the negotiation occurs: (i) Progressively (for each derived
bridge a negotiation procedure occurs in that moment) and (ii) At the end (each bridge is negotiated at the end of the bridging phase). A variation of these two approaches is possible, negotiating a group of bridges instead of individual bridges.

These two approaches are mutually contradictory with respect to advantages and disadvantages, but it is possible to combine both approaches, coping with each other's disadvantages while maintaining the advantages.

3. Semantic Bridges

Semantic associations are categorized according to (i) Entity type, (ii) Cardinality and (iii) Scope. This section describes several issues in dealing with these dimensions.

The *entity type* dimension refers to the type of elements being associated in the bridge, which are of a single type. Typically there are 4 types of entities being described:

- Concept corresponds to compound objects of attributes, relations and axioms. It corresponds to the OO class entity (RDF, 1999);
- Relation corresponds to a connection between two concepts, the subject and the object forming a triple (RDF, 1999). Each concept has its own identification, scope, etc.;
- Attribute corresponds to a connection between one concept (the subject) and an atomic value (the object) (RDF, 1999). The attribute has no identity, and its scope is the concept it belongs to;
- Extensional Patterns intend to describe data source instances. This arises from the fact that ontologies describe entities from different perspectives and details. Hence, sometimes it is necessary to expand the ontology characterization to meet the other ontology specification. The case from Figure 1 is a good example of this requirement.

Some other type of entities like statements, rules, functions, etc. are allowed in some ontology languages, which may appear in some ontologies. One may suggest these entity types should also be bridged. However ontology alignment does not intend to provide alignment between ontologies but between data sources described according to ontologies. I.e., such elements are part of the ontology and not of the data source; thus, the alignment process should not align these elements. On the other hand, these elements should be considered in the alignment process, once they describe information about the entities in the ontology.

The *cardinality* dimension relates to the number of objects intervening in the bridge from both ontologies, ranging from 1:1 to m:n. Although it is theoretically possible, we did not find any real case where a m:n bridge was necessary that could not be substituted by 1:n and n:1 bridges whether entity type the bridge respects to.

The *scope* dimension relates to the bridges inter-relation and application, similar to the OO modeling perspective (composition, inheritance, abstraction, etc.). We suggest two modeling approaches that corresponds to three constructs:

- Composition allows specifying a bridge as being composed by several others. For example, the concept bridges presented in Figure 2 are both composed of other bridges. Functionally it means that, to complete the outer bridge, the inner bridges should also be
Inheritance allows specifying an hierarchy of bridges, from super-classes to sub-classes. Functionally it implies that the super-class bridges should be called by the sub-class and executed prior to itself. Abstraction is the third construct and represents a variation of the type of the super-classes. When this attribute is set, the specified bridge should be used only as super-class of another, and should not be executed independently. This situation is described in Figure 2 in the relation between the two concept bridges. Notice that the first bridge is stated as abstract, once it is considered that each instance of Staff should always belong to either Employee or Manager.

Figure 2 – Ontology bridging example

The n-1RelationBridge defined in the example from Figure 2 represents a bridge between two relations (shaded) in the source ontology and one relation in target ontology. It may be understood as:

```javascript
var o1.Staff s="Jim", "Juice", "Orange 23, 4234 orange grove", 122,323, "#Food";
var o2.Employee e;
e.name=concat(s.firstName, s.familyName);
...
var o1.Department d=s.worksIn;
var o1.Staff m=d.isManagedBy;
e.hasManager=m;
...
```

ontology 1

```
c_1 \rightarrow r \rightarrow c_2
map
```

ontology 2

```
c_1 \rightarrow r \rightarrow c_2
map
```

Figure 3 – Relation bridge (validity conditions)
Notice that a relation bridge makes sense only when three conditions are verified (Figure 3):

- The sequence of the relations are set and maintained;
- There exists a bridge between the concept holding the first relations from both ontologies;
- There exists a bridge between the concepts that play the object role in the last relation of the bridge.

In addition to previous constructs, we believe that it is necessary to specify a formalism to represent constraint rules associated with the bridge. This formalism allows us to specify several bridges for the same entities, but prevents execution according to such a constraint rule. In the example from Figure 3, we could define a bridge C similar to bridge B, but associating Manager instead of Employee. A constraint rule would be associated, constraining it to be executed only if the Staff instance is referred as the object in at least one instance of the Department.isManagedBy. In that case a Manager instance is created instead of Employee.

4. Summary and Future Work

We suggested a decomposition of a traditional ontology merging process, but we suggest the independence of specification and execution of semantic bridges. The proposed strategy permits aligning heterogeneous data sources or heterogeneous ontologies in an inexpensive and effective way. Transformation relations are specified between source and target, dynamically and independently of contents, permitting an on-demand fast query mechanism. This strategy was first introduced and better justified in (Silva and Rocha, 2002). Accordingly, we suggest the existence of a representation formalism that would allow us to exchange correctly and unambiguously the specified semantic bridges. We suggest describing such formalism in a meta-ontology, which would allow to represent and exchange them as an instantiation. Hence, this meta-ontology would serve simultaneously as a representation and validation formalism.

Currently we are focused on formalizing the semantic bridge conceptualizations described in 0, and in specifying this meta-ontology. The real semantic bridges cases allow us to test such a language, and especially, test the usefulness and completion of the one composed of the described constructs for the scope dimension. This work is also progressing by automating as much as possible the association phase.

Acknowledgment: This work is partially supported by the Portuguese MCT-FCT project POCIT/2001/GES/41830.

References


Sintek and Decker, (2001); M. Sintek and S. Decker; TRIPLE - An RDF Query, Inference, and Transformation Language; Deductive Databases and Knowledge Management (DDLP'2001); Japan; October 2001.


W3C, (2002); World Wide Web Consortium; [www.w3.org]; accessed 22/03/2002.
Giliola Negrini and Patrizia Zozi
Consiglio Nazionale delle Ricerche - ISRDS, Roma, Italy

Ontological Analysis of Literary Work of Art

Abstract: Ontological structures can aid the understanding and modelling of works of art. Ontology of the aesthetic object, and particularly of the literary work, has been analysed by Hartmann and Ingarden. Application of Dahlberg's ontical 'systematifier' model enabled us to organize the entire structure of the Thesaurus of Italian Literature, and to highlight a number of significant aspects of the literary work. After describing the conclusions arising from the experience of compiling the thesaurus, the paper briefly outlines Hartmann's and Ingarden's theories of levels and seeks to identify commonalities between the ontological analysis of the two theories and the conclusions of the thesaurus.

1. Introduction

The interest in the studies referred to in this paper was prompted by Roberto Poli's article 'Levels' (Poli, 1998), where the author emphasises the difference between levels of reality and levels of description and specifies that levels of reality have a strictly ontological valence, while those of description have a strictly epistemological one. On analysing the theories of three groups of authors, Poli points out that the problem of levels is present in their works and that a number of reciprocal influences emerge.

Following Poli's reasoning, we turned our attention to the theories of the phenomenologists Nicolai Hartmann and Roman Ingarden, who studied the aesthetic object, and the literary work in particular, from an ontological point of view. Their theories were of close concern to us because we had organised the Thesaurus of Italian Literature and set out reflections which seemed to be a milestone in the ontological results arising from the practical structuring of literary terms. Application of the ontological 'systematifier' model enabled us to organize the entire structure. It also enabled us to show that the literary work is a complex structure, and to highlight some of its distinctive features.

In what follows, we shall first present the results of our work on the thesaurus and discuss the principle of levels set out in Hartmann's and Ingarden's theories. We shall then show a number of features shared by the two theories and compare them with our own conclusions from our structuring work.

2. The conception of the literary work that emerged from our work on the Thesaurus

The Thesaurus of Italian Literature is based on an ordering of the subjects indexed by LI.AB (Letteratura Italiana. Aggiornamento Bibliografico), a current bibliography on what has been published on Italian literature from its origins to the present time. In order to organize the subjects, we used the 'systematifier', a set of categories devised by I. Dahlberg to organize the plan of a conceptual system (Dahlberg, 1991, p.110). These categories provided the framework for the
organization of our literary terms, the first three of them in particular: what we
called the ‘core categories’ and which enabled us to draw a number of key
conclusions on ‘what literature is’.

The core categories were the following: (a) theory and principles related to
the domain’s object; (b) object’s description; (c) activities and processes related to
the object.

In order to apply them to the field of literature we had to identify its
ontological features. Since the concept literature refers to a set of literary works,
and the type of literature is characterized by the kind of work examined, we
identified the literary work as object of our analysis. Consequently it was to the
nature and properties of the literary work that we had to refer.

Our analysis showed that the literary work is a complex phenomenon which
involves numerous characteristics, many of which stand in opposition to each other.
This opposition emerges in following aspects: materiality and immateriality;
determinacy and indeterminacy; truth and non-truth. Materiality represents the
‘external dimension’ of the work, while all other characteristics concern its ‘internal
dimension’ or content. We called this last the ‘literary fact’, an artificial construct
used to represent an event that occurs at a particular moment and in a particular
place, and to which the author transfers his/her creative image. The literary fact
comprises not only the sentiments, passions and ideals that have inspired the author
but also the way in which they have been expressed.

Materiality and literary fact identify the literary work as object. Consequently
description of work and types of works should be attributed to the
category b) ‘literary work as object’. In organizing these concepts our attention was
devoted to the literary fact. We could establish that the type of work is the way in
which the literary fact is presented, as in the following examples:

<table>
<thead>
<tr>
<th>according to the form</th>
<th>according to content</th>
<th>according to mode of production</th>
<th>according to final production</th>
</tr>
</thead>
<tbody>
<tr>
<td>anthology</td>
<td>almanac</td>
<td>Imitation</td>
<td>unfinished work</td>
</tr>
<tr>
<td>dialogue</td>
<td>bestiary</td>
<td>original work</td>
<td>unpublished work</td>
</tr>
<tr>
<td>play</td>
<td>biography</td>
<td>Plagiarism ...</td>
<td>first edition ...</td>
</tr>
<tr>
<td>collection of letters</td>
<td>chronicle</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poetry...</td>
<td>diary ...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tab.1.- Typology of literary work

The category a) ‘theory of the literary work’ is analysis of the latter’s
structural constituent foundations as well as of the processes and methods used by
literary criticism to identify its elements. The subjects falling within this category
were grouped according to whether they concerned theory of the literary fact,
theory of criticism, theory of translation.

The creation of a literary work by the author is the process that produces the
literary fact; interpretation is the process that seeks to clarify that fact. Like the
reader, also the critic interprets the literary fact, and the interpretation furnished is
'value added' for comprehension of the literary fact. The category c) consisted of criticism as action taken on a literary work. It was decided that all of the intellectual activity carried out with a view to acquiring knowledge about the literary fact should be regarded as constituting the 'process concerning the object'. This focuses on the action and procedure of interpreting the content of a work in order to grasp the author's creative image. On the basis of these principles it was possible to identify further categories and consequently classify all the other subjects.

In our reflections it was emerged that each interpretation may be questioned and subjected to further criticism and a literary work can never be entirely brought to light.

3. The thought of Nicolai Hartman

Hartmann did not study the literary work as an ontological object, but rather the work of art in general. However, his detailed analysis of descriptive and non-descriptive arts highlights the features that characterize the poetic work, distinguishing it from others. It is to this analysis that we refer in this section. The fact that Hartmann frequently makes explicit reference also to not exclusively poetic literary works enables his theory to be extended to the literary work in general, and thus to be compared with Ingarden's thought.

A poetic work is an object in which it is a concrete work created by a writer, and a representation. Aesthetics seek to reveal the knowledge concealed in this representation.

A work-object is shaped from physical-material elements: language, words and the act of writing make the work real and shape its materiality (Materie). These elements serve to express a content, a theme (Stoff), in which the sentiments, pains and destinies which the author wishes to communicate are present. The representation emerges from this content. The structure of the content is of particular importance in Hartmann's theory.

Two closely connected levels can be distinguished in a poetic work: the foreground (Vordergrund), which is real because it is that of materiality; and the background (Hintergrund), which arises from the former, concerns the content (Gehalt) and its structure, and is ideal. A poetic work is ontologically characterized by this opposition. The perceiving subject intervenes but remains extraneous to the stratification.

The words that describe settings, characters, actions and feelings create an apparent world in which the reader participates as if it were real. The author of the poetic work indirectly expresses his/her ideals or psychological states, and through the personages created by his/her imagination transmits his/her life-experiences more or less overtly. The concepts conveyed by the words provide the reader with a vision (Anschauung) which may be limited to, or contrary to, the covert one that inspired the author. As deeper-lying contents are uncovered, they reveal forms, values and destinies which are not described materially. These visions generate appearance relationships (Erscheinungverhältnis).

Compared with those of the other arts, a work of poetry has the most complex content and the highest level of ideality in comparison to the semantic poverty of the words. A poetic work, therefore, does not have one simple background layer but a whole interconnected series of them. The background, in fact, breaks up into the fantastic profundities of the ideas, not immediately, but
mediately through other non-real and aesthetic layers.

From this one may conclude that a poetic work, but also a play, a novel or an epic, presents:

- a foreground and a background according to it 'manner of being';
- several layers according to the 'overall structure of the content' or its 'inner form' (Hartmann identifies seven such layers: 1966, p. 183).

The essence of the poetic work as an aesthetic object resides in the appearance relationship. The ability of Hartmann's aesthetic inquiry to make every layer of the background visible is its key achievement.

4. The thought of Roman Ingarden

Ingarden examined the ontology of the literary work in his essay 'Das literarische Kunstwerk', in which he described it as an intentional object requiring both real and ideal elements.

A literary work requires a real object because it is a material entity independent of either the writer or the reader; and at the same time it is a representation requiring ideal objects as well. The set of sentences constructed by the author – which results from his/her subjective operation of representing ideal contents originating in his/her experience – constitute the work. By means of an intentional subjective operation, the author actualizes elements (words) which signify corresponding ideal contents, so that the sentences are not ideas but formations (Gestaltungen), pure objects of the author's representation. The work has its identity and uniqueness, and a heteronymous essence. Ontologically, it is something different from the mind of the author who creates it, and from the mind of the reader who interprets it and draws his/her own image from it (concretization).

A literary work is a single entity but it comprises several heterogeneous strata and layers, each of which has its own distinctive features and role. The most important levels are the following.

1. Stratum of text and of sound forms. This is the material of the work, its external form. It concerns the choice of the words and sentences that convey meaning and represent the constitutive part of the literary work on which all the other levels depend.

2. Stratum of meaning. This stratum has several layers.

2.1 Layer of meaning units. This first layer comprises the meaning of words and sentences. It does not have autonomous essence; its origin and essence depend on conscious subjective operations performed by the author. Through his/her choice of words the author partially actualizes his/her ideas by formulating part of the material and formal content. The composition of the sentence introduces a meaning additional to those of the individual words, and it represents an intentional and functional unity. As a purely intentional expression it has a dual structure, which comprises the fact described and, more broadly, the ideal content represented. Characteristic of the literary work are quasi-judgements (Quasi-Urteil), statements that are only apparently true and which stand in opposition to true judgements (Urteil), which are statements which meet the truth criterion typical of the scientific work.

2.2. Layer of represented objectivities. This next stage in the acquisition of meaning enables the production of an image broader than that described. The
represented objectivities include, in fact, both ideal contents intentionally represented by the author by means of words and sentences, and whatever is acquired through properties, not expressed but possessed, of the things persons and facts cited. These properties constitute a background whose boundaries cannot be determined. This layer seeks to identify which level of reality is appropriate for the quasi-judgements by taking account of the place and time of representation. The presence of these phenomena accounts for the impossibility of achieving complete understanding of the work, which will only ever appear partially: hence the presence of points of ‘indeterminacy’ in the represented objectivity.

2.3. Layer of schematized images. This expresses the capacity of the work to bring out schemes of images that the author wants to exhibit: figures used in historical language, metaphors and similes. Schematized images confer a visibility wanted in a particular way on the object represented. The facts present in the images do not concern what the object is (its essence) but what it seems. The images have their own aesthetic values which are part of the work’s general aesthetic value.

5. Features shared by the two theories

The spread of the two theories with their similarities and differences led to heated polemic between their propounders. From a scientific point of view, however, convergence of thought is “regarded as highly significant. When several researchers working independently obtain similar results, it is reasonable to assume that there is some accuracy in their theories” (Poli, 1998, p. 206). Therefore, in order to identify the ontological characteristics of the literary work, we shall seek to show the most evident similarities between the two theories.

1. Conception of a literary work. A literary work is an object: that is, something in itself, something that is created and which once created no longer depends on its creator. It is representation as such: only subsequently does what has been represented emerge. Neither Hartmann nor Ingarden deny the presence of either the author, who transmits his/her feelings and thoughts by creating, or that of the perceiving subject. But the work exists as an object with representative content independently of the psychological aspect of subjects. This also applies to the author who has created it, for the existence of the work began at the moment when the author finished creating it. The work has heteronymous nature.

1. **Levels.** Both theories are based on the concept of levels and of a plurality of strata and layers as constituting the structure of a literary work. They express an autonomous and definitive reality, within the unity of which it is possible to distinguish the elements that constitute it.

2. **Materiality.** Hartmann’s *Vordergrund* corresponds to Ingarden’s stratum of the text and sound forms: it is this that gives form to the object and transmits the content of the work.

3. **The concept of background.** The two philosophers often use terms whose meanings sometimes coincide and sometimes differ. The term background, which is of particular importance in both theories, does not denote the same concept. For Hartmann it is the stratum counterposed to materiality and it possesses its own structure. In Ingarden’s represented objectivation, it is the ‘obscure part’ which remains outside the ‘light cone’ of the meaning of the words, and belongs to
a particular layer.

4. The poverty of word content. In different ways, both authors emphasise the limitedness of the meaning of words, and of the literary text in general, compared to the deeper-lying and ideal images that are present.

5. Object and subject. Because a literary work has a very high level of ideality, represented by the author and grasped by the reader, the theory which attributed subjectivity to the literary work used to prevail. Hartmann and Ingarden’s ontological interpretation rejects this theory by justifying the existence of both an object, in which the ideality is represented, and of subjects, one which it represents and others that interpret it. The fact that the reader is able to see and interpret does not gainsay the presence of the intentional and functional object.

6. Comparison between the theories and the conclusions of the thesaurus

We shall finally seek to show the similarities between Hartmann’s and Ingarden’s theories and the conclusions drawn from our work on the thesaurus.

Our conception of a literary work as an object that express the literary fact corresponds to that present in both theories. The literary fact represents the meaning, in fact the literary fact ideally comprises everything that concerns the content of the work, its spiritual and aesthetic values but also the forms of the work, because these too are an expression of the author. The external dimension represents the materiality. Nevertheless the material in our conception regarded work as final product including its publication, support, etc., materiality of the theories identifies what is written by the author.

For this reason the external dimension does not emerge in the thesaurus structure. We focused our attention only on the literary fact which suggested solutions to structure problems. Terms relative to grammatical analysis and linguistic analysis, to verse theory and prose theory, did not refer to the text as the materiality of the work but rather to expressive values. Terms like unfinished work, unpublished work, first edition belonging to the category ‘object’ and structured ‘according to final production’ concerned the way in which the literary fact was presented (see section 2). As regards terms referring to constitutive parts of a work’s form (introduction, preface, prologue, text) our point of view was very closed to that of Ingarden’s. He declares the non-existence of the beginning, parts and end of a literary work: elements which he considers to be temporal spaces. Ingarden sees the ordering into parts of a work as a succession of phases in the unfolding of the story, or in other words, as a sequence with which to build the structure of the work.

Terms relating to other categories of the thesaurus evidence a deeper level of content. These terms are those which concern the main theme of the work, the elements highlighted by criticism, features of the literary movement and phenomenon, the physical person/literary work relationship (author, mediator, reader).

The ‘indeterminacy’ of the literary work posited by Ingarden was also borne in mind when compiling the thesaurus. The effort to interpret the literary fact is continuous. Every form of interpretation brings to light diverse aspects of the literary fact, which in their turn are subject to the further study that drives the search for interpretation. It has been pointed out that interpretation proceeds both horizontally, with perception of the work’s content by several readers, and
vertically when increasingly more extensive images overlap with previous ones. A crucial role in this latter form of interpretation is played by literary criticism, to which the thesaurus devotes particular attention.

We conclude with two observations: 1) the foregoing discussion is only an outline of the many stimuli for reflection that arise from the thought of Hartmann and Ingarden. Their works are a ‘mine’ of knowledge about not only works of art but also every form of writing in general; 2) the two theories enabled us to develop the ontological knowledge yielded by application of the ‘systematifier’ model. We believe that seeking to understand the profundity of the two theories, grasping their validity, and showing their topicality, having had the simple experience of classifying indexed terms, demonstrates the robustness of the ontological model used, the principles of which suggested the concept of an ontology of the literary work.

Notes
1. The term ‘level’ refers to both ‘layer’ and ‘stratum’. Layer identifies overforming relationship, where overforming means that every category can constitute the ‘matter’ of a higher category. Stratum identifies building-above relationship, where building-above denotes that the higher stratum requires the lower one only as its external basis of existential support, but not as matter to be supraformed. (Poli, 2001, p.124)

References


Hartmann, Nicolai. (1966). Ästhetik. Berlin: de Gruyter


Abstract: The objective of the paper is to analyse the information (communication) process of fiction in order to increase our understanding of the different actors and factors involved in information processing. The methodology is that of the grounded theory, where previous studies of information process and fiction content representation are compared with the results of an empirical study in which library patrons and library professionals were asked to index and abstract five different kinds of novels.

1. Introduction

The mechanical model of information transfer was defined during the 1940’s for electronic communication tools, especially for the telephone (Shannon & Weaver 1975, p.7 and 33-34). In that model the communication was seen as a one-way process, i.e. the message delivered from the transmitter to the receiver should be as similar as possible.

Although this model is still widely used as our common perception of the information process, it has been criticized as being too simplistic. (See e.g. Foskett 1996, 5-7.) Eco has defined this as the model of the elementary structure of communication, since it describes only the physical stream of the communication and does not take into consideration the actual transition of meanings and the different kinds of interpretations that can take place in a communication between two or several persons. (Eco 1977, 32-33.)

In addition to the physical information transfer process, one can discern a parallel cognitive process (the actual communication process), where the knowledge of the actors is transmitted within the processed information and thus the sender tries to influence how the receiver perceives the information and vice versa (Järvelin 1995, 4-5). The model of the elementary structure of communication lacks also other crucial aspects of human information transfer and processing, for example the context and its effect on the communication as well as the effect of feedback during the communication (see e.g. Kovala 2001, p. 11, Riessman 1993, p.14).

2. Theoretical Approach, Research Questions, Data and Methods

The basic theoretical approach used in this study was qualitative, specifically the use of the grounded theory approach (see. e.g. Strauss & Corbin 1990). In the analysis of the obtained empirical material, both qualitative (mainly content analysis) and quantitative methods (mainly statistics) were used. This so-called triangulative approach was chosen to confirm the findings of the study (see e.g. Strauss & Corbin 1990, 18-19). The analysis of the indexing was carried out mainly with the aid of statistical and mathematical methods (see Saarti 2002), the
analysis of the abstracts predominantly with the aid of content analysis (see Saarti 2000).

Five novels from different genre categories were chosen for the experiment. A total of 30 people were chosen as indexers and abstractors for these novels. They were selected from five different Finnish public libraries: three patrons as well as three library professionals from each library. In order to ensure the comparison between different test subjects, they were all required to read the same novels. Then they were invited as a group for a discussion and asked to complete a questionnaire. In the group session, they all wrote abstracts and indexed the above novels. The indexing was done with Kaunokki – the Finnish thesaurus of fiction (see Saarti, 1999), the use of which was demonstrated to the test subjects.

The results from the analysis of the content description were analysed together with the results derived from the previous studies of the fiction knowledge organisation (see e.g. Beghtol 1994, Pejtersen & Austin 1983 and 1984), as well as with the analysis of the fiction communication (see chapter 3). The comparison was done in order to analyse the effect of the fiction information process on the knowledge organization of fiction.

3. Information Process of Fiction and its Main Actors

The main participants in the information process of fiction are: the work of art, its creator (i.e. the writer), the reader and the social-historical environment where the publishing and reception takes place (see fig. 1).

Due to the nature of fictional works, the reception of the work of art cannot be considered as complete, unless all of the above mentioned participants are involved in the process. The role of the writer is to write works of art – novels, short stories, poems, plays etc. – that are to be published. The role of the work of art is to be a medium by which the artist can communicate with his/her audience. Also the work of art has its own, autonomous life: after the book has been published, the writer’s only subsequent role as a readers, i.e. an interpreter of the work.

The role of the reader is that of the interpreter of a work of art. The interpretation as well as the creation of a work of art takes place in a social-historical context that defines the language used and its means of artistic expression. Without this common language, no communication between readers and writers would occur.

Fictional communication is also typically bi-faceted. On one hand it consists of factual meanings, i.e. references to actual happenings, historical and geographical facts etc (see e.g. Ranta 1991, 20-23) and on the other, it has an aesthetic function and is thus based on the individual interpretation and reception of the work of art.
4. The effect of the Fiction Information Process on the Actual Content Description

The content analysis of the abstracts revealed that about 75% of the 3206 different elements found in the abstracts described factual aspects of the novels (e.g. themes, settings, characters). Other categories (and their percentages) were: describing the novel structure (11.9%); describing the subjective reading experiences (5.5%); describing the novel historical setting (0.9%); references to the author (1.6%), and criticism and evaluation of the novel (5.2%) (see also Saarti 2000).
On the other hand the statistical facet analysis of the keywords used showed, that about half of the keywords used described thematic aspects of the novels. (The thirty test persons in this study made use of a total of 632 different keywords in indexing the five novels.) The next most popular category, used by about 20 %, described characters in the novels. Keywords describing settings and genres were both present at a rate of about 10 %. The least frequently used keywords were in the categories of time (5 %) and others (1 %, e.g. classics, translations) (see also Saarti 1999).

Fig 2. A broad model for a search and retrieval system for fiction (Saarti 2000, p.219).
Thus one can draw the conclusion that all aspects of the fictional communication process are present also in the actual content description of the novels. However, as found in this study, it is also quite evident that the most important focus is on the actual content of the novels – the other aspects have only a minor and supporting role.

5. Conclusions and Discussion

Although the corpus of this study was limited, certain conclusions can be drawn. Different readers can be seen to use radically different elements when they are describing a novel’s contents and the aesthetic experience of reading it. Also all the important participants of the information process of fiction are found in the abstracts, i.e. work, author, reader and the cultural-historical context. The description of fiction’s totality (its aboutness) must include the entire process of fiction and its reception, not only the texts and their contents. On the other hand, these latter two aspects can be seen as the most crucial factors in content representation of fiction. Furthermore there is already some evidence (e.g. from the internet bookstores) that also subjective aspects are important in fiction information systems. This, of course, is one topic that needs further study.

One clear result of studies on indexing and abstracting of fictional works is the impact that the interpretation of the work of art and the context of its reception have on the content description of the works. This could also be clearly noted in this study when analysing the abstracts that were produced. These interpretative and contextual aspects of content description are a subject for further study, not only in fictional works but also in scientific material.

It is apparent that fiction search and retrieval systems have to be multifaceted in order to meet the varying needs of different users. With the aid of this kind of system (see fig. 2.), one can document holistically the different aspects of the meaning of fiction, i.e. what the fiction is about. The building of these kinds of systems has started, but there is still much work to be done.

Acknowledgements

I am grateful to Dr Ewen MacDonald for revising the English.

References


Constitution of the Scientific Domain of Information Science

Abstract: Objectives: To delimit the epistemological parameters of Information Science on the basis of the terminology it employs. Starting from the hypothesis according to which the "central" terms should be characteristic of, or specific to, the area, making up the hard core of Information Science, while the "peripheral" terms would have originated from other knowledge areas, with which Information Science interfaces. The study is justified in that Information Science is a scientific domain in development, which does not present consensus as to its object and does not clearly define its boundaries — or interfaces — with other areas of knowledge.

Methodology: The terminology utilized by Information Science borrows terms from the areas of Logic, Administration, Linguistics, Computer Science, Sociology, Communications, Cognitive Sciences, as well as the terms coming from Librarianship. Research was therefore effected on the basis of the following terminological procedures:

a) delimitation of the domain: initial set of terms
b) selection of bibliography
c) development and filling in of research forms
d) development and filling in of terminological forms
e) consolidation of terms (vocabulary)
f) elaboration of definitions.

By means of these procedures, the domain's "central" and "peripheral" terms were detected.

Results: The analysis of terminology permits the following provisional conclusions: the central terms, and therefore those specific to Information Science, come in large measure from the area of Librarianship and today constitute the field of "Information Organization and Representation", while terms originating in other areas were not, in the majority of cases, submitted to processes of adaptation, sometimes maintaining their original meanings, but more frequently the terms were adopted while their original conceptual structure was abandoned.

The paradox may be stated thus:
- The interdisciplinary character of Information Science is clearly upheld in the majority of discourse, or practically consensual, from the 90's;
- Information Science, taken as area of management of recorded memory, emphasizes its importance on the modern and post-modern stage of the social sciences, but rarely proposes itself as a science that discusses principles or laws, in function of its history: the history of Information Science is systematically suppressed in the name of its underlying modernity and technology, or on the other hand assimilated into the history of Librarianship;
- Although Information Science, as an applied social science, does not coherently mobilize the characteristics of a fully fledged science, it proclaims itself as an area that aspires to intervene practically in society through the optimization of the information flow and the democratization of access to information;
- In spite of the almost exclusively argumentative investment in the social and scientific character of Information Science, the real conceptual investments were made in order to
optimize the internal procedures of the area, in their majority originating in Librarianship and Documentation;
- The discussion of the procedures specific to the area allows the recovery of the communicational dimension of information representation, but this discussion is still peripheral and viewed with doubt or suspicion by many authors. The communicational dimension of information representation is absorbed, and rarely discussed as a consequence of the logico-linguistic dimension of information production and organization.

The identification of Information Science as a knowledge domain implies the development of an individual terminology, critically discussed and systematized concepts: that is what we offer as a product of the present study.

1. Object and objectives

It is undeniable that the domains of communication and of information have advanced during the last century, but the constitution of the correlated scientific fields faces obstacles of various kinds. Among the most important of these is the lack of consolidation of the conceptual corpus used in the analytic and discursive production of these areas. Thus, when we refer to them, although there may be agreement that we are dealing with important practices, fundamental for the transformation of the world, as they are responsible for a culture which subverts the traditional typologies of the manner of knowing, we do not recognize the thought which constitutes the fields in question.

In the specific case of the Science of Information, its understanding as a scientific field depends in large measure on a nominalism, that is, on an instituted meaning form – the denomination Information Science itself –, whose reference, most often, refers back to the inter- or multi-disciplinary universe, that presents a not very distinctive character. In fact, such a situation does not resist enquiry as to what is the specific object of the field, which are the principal problems to be faced, its hypotheses and the nature of its procedures. Imprecision, in this case, results from the fact that one may readily replace the notion of interdisciplinarity with that of union or junction. Might not Information Science be a union of procedures from different disciplines with the single objective of developing a practice?

The relation between man and information is certainly older than the first library. But it is starting from this facility that specific social modes of reception of recorded knowledge were instituted: the library was a place of retreat, the place for rare works, the place of selection of order, that of the collection. Through this facility and its variations were established orders and forms of socialization and of use of stored information. Naudé said it in 1644 (Naudé, 1876) and Otlet in 1934 (Otlet, 1934). In spite of all the conceptual settling that this rich experience could have provoked, the disciplines involved ended up proposing practical notions, passing wide of a consolidated theoretical network. From the point of view of the strict link they maintain with recorded knowledge and its organization for socialization, Librarianship, Documentation and Information Science have presented too loose a terminology, sometimes stemming from terms borrowed from different disciplines, such as sociology, economics, computer science, psychology, etc. This mixed composition brings consequences. The terminological imprecision provokes not only impasse but also strongly collaborates toward theoretical retardation. As the disciplines take up their spaces in the universities, courses multiply, imposing manners of circulation similar to the other graduation courses,
the terminological problems pass into the foreground.

Starting from this context, the present project, still in development, proposes the question about the epistemological status of Information Science, observed from the terminological vocabulary it employs.

2. Methodology

It is a recognized fact that the denominations serve as reference for the determination of the vocabulary of a specialized field, that is, of the conjunction of significant forms that respond for the particular concepts from which the fields of knowledge are constituted (Cabré, 1993). In this way are integrated the special vocabulary, the concepts relative to the objects, processes and methods that permit the development of the investigation and the production of knowledge. As the concepts do not result from absolutely arbitrary conventions or from individual preferences, but from the relations between their characteristics and their pertinence to the domain, as starting point, a hypothetical and provisional configuration of the domain was determined. It is considered that Information Science is the discipline that operates with questions relative to:

- **Production**: identification of the explanatory codes of the contents recorded in the form of information (result of the operations on the recorded contents that appears in the form of socialized contents), articulation between the technological mechanisms and the production of information;
- **Circulation of the information**: social insertion of the information;
- **Information consumption**: conditions for reception of information, places, facilities and users; sociological, political and economic dimension of the informational activities.

The first stage in the investigation was dedicated to the identification of the limits of the area and selection of a set of denominations that take account of these limits.

The survey and selection of the bibliography, to constitute the terminological documentation involved:

A bibliographical survey in the LISA (Library and Information Science Abstracts) covering the period from 1972 to 2001. This first survey, deliberately broad-based, was operationalized from the start point of two descriptors: Information Science and terminology, detecting 90 bibliographic references in the research. This list was submitted to a selection, guided by the following parameters:

- texts which, in their abstracts, demonstrated conceptual and historical concern in relation to the area;
- availability of the texts;
- distribution of the texts throughout the period, a high incidence of texts being noted in the 70's and 80's;
- the co-presence of articles from periodicals and books;
- the language of the texts, discarding some references in Japanese, Chinese and Russian;
- the inclusion of authors considered “classic” for the history of the area, such as Buckland, Hjørland, Lancaster, Line, Liton, Rayward, Saracevic, Shera, Vakkari, Vickery, Wellish, Wersig e Yeuxiao. At this stage we were not concerned about including Brazilian authors, who were later added to the survey.
We worked, then, in the end with 29 texts.

Reference works – the dictionaries and encyclopaedias of the area were included, covering the period of 1973 to 1997. Special attention was paid to the “Dictionnaire Encyclopédique de l’information et de la documentation”, organized by Serge Cacaly, and edited in 1997, which was compared with “Terminology of documentation” organized by Gernot Wersig and Ulrich Neveling for UNESCO, in 1976.

Methodological documentation – the terminological standards ISO 704, ISO 1087 and ISO 5127 underpinned the methodology of the investigation.

3. Notional identification procedures

From the bibliography 62 terms were identified in the condition of recurrent terms, and thus essential to the analysis of the notional field of Information Science. For these terms definitions were then gathered and a synthesis elaborated.

At the end of this stage of the research a test was carried out of the categorization of the notions of the area, according to their status in relation to the area of Information Science:
- notions of the area, laid down over time;
- notions of the area, semi-elaborated;
- notions of the common empirical experience;
- confounded notions.

Notions of the area, laid down over time – in this category are included the notions that designate the classic procedures and components of Librarianship, such as “classification”, “library”, “abstract”, “documentary analysis” and “invisible college”. However, although these notions present stability in their concept and tend towards univocity, it is interesting to note that these, in the main, are described by the procedures or by the components that determine their reach, but rarely by their pertinence in relation to the area. Thus “classification” for example, is defined by the procedure and by the instruments used, but not by its function in relation to the area of librarianship or Information Science: the “organization of information” is not, in other words, described as a means but as an end in itself.

Notions of the area, semi-elaborated – these notions, occurring with frequency in the survey, point to terms whose concepts still show great variation, in that the different approaches do not appear as effectively distinctive. In the bibliography, these notions are very often presented having various meanings, or approaches, but the differentiation does not imply schools of thought or different traditions, but a simple coexistence of a diversity of opinions. In this category may be included such notions as “need for information” (which does not discuss the difference between needs, demands and wants), “information science” (which accepts very different amplitudes when the term “information” is not attributed a definition specific to the area, and which differentiates it from the areas of communication or from empirical experience), “analysis”, “search strategy” and “information retrieval”.

Notions of the common empirical experience – Different notions – central to the area – are defined in a totally empirical form and therefore do not configure a terminology specific to the area. Thus, for example, the term “book” is described as a set of printed pages, linked together, containing at least 49 pages. The function of the book for the area, that is, as an information holder, is not noted, there being no
distinctive concepts of the book for the area of Information Science or of Publishing. When the definitions of “book” go beyond its materiality, they inscribe it in the list of sacred objects, worthy of worship, and the vagueness of the term remains as far as its insertion in the area is concerned. The notion “information transference” gives occasion to paraphrases but no distinct delimitation in relation to the area. The notion “subject”, in its turn, gives rise to definitions inserted in the area (principal theme of a book) or others, drawn from empirical experience (concept which introduces, justifies, proves or amplifies the principal theme of a work). The notion of “user”, very frequent in the bibliography of the area, does not have a distinctive definition, as it is considered synonymous with “reader” or “client of the library”.

Confounded notions – in this category we include terms “imported” from other areas of knowledge and not submitted to an adaptation to the area of Information Science. The notion “data base” exemplifies this category. The phenomenon of the terminological importation, very frequent in Information Science, indicates an area that maintains very close interfaces with a diversity of other areas of knowledge, but also permits reflections as to the reasons for the absence of “exports” of Information Science terms to other areas.

4. Final Considerations

In general, linguistic expressions are used in different ways. Observed through classical logic, such expressions have the function of letting us know the proposition, the idea, that is, the objective reality to which the concept refers, or allowing one to attribute it a judgement. However, in the ambit of the Human and Social Sciences, the ambiguity and the equivocation that distance us from the expression of clear and distinct ideas, do not result from an insufficiency of the language. One cannot eliminate from them the contingency the historical process and the social reality, in general the possibility of diverse interpretations, in majority based on experience.

In this context, characteristic of the whole controversy, the fundamental condition of logic – univocity – cannot be obtained. Thus the non-existence of an equal and identical range from beginning to end of the arguments is identified: the demonstration of the concept is not finalized, whether because all the criteria of applicability are not identified, or whether because the notions are still empirical.

In general, an elaborated conceptual system is not identified in the area. The conclusions are obtained by interpretation, that is, it is possible to attribute a hitherto unforeseen meaning to the notions. This is determined by the very nature of the information.

It is customary to consider as foundation texts of a scientific area, those which present a theoretical-practical base, supported simultaneously in a general and specific knowledge, in such a way as to propitiate the internal advance of the area. On the other hand, what is found in Information Science, are texts whose ends pursue self-legitimization: they attribute to the area an imperative social significance, present the methodological guidelines in a conclusive form, leading to a positive criticism that limits the proposal of controversial topics. (Day, 2000).

The domain of Information Science is not constituted only by referents with which it is concerned but principally by the focal points and instruments that associate these referents with situations that determine the field of production and
circulation of informational goods and services. This being so, the area finds itself between the totality of a compartmentalized discipline – derived from the experience of cultural life – and a project founded in partial sciences disconnected from one another. It is organized initially by means of a scheme of concepts linked to distinct domains and gains the specificity of situational discipline, when it proposes to represent the world in a certain way and intervene in it through certain social products and equipment.

Notes
1 Research developed with the support of the Brazilian National Council for Scientific and Technological Development - CNPq.

References
Anita Coleman  
University of Arizona, Tucson, USA

A Classification of Models

Abstract: Libraries, bibliographic databases, and the WWW/Internet do not provide access to a coherent, unified class of works known as models. This paper presents a framework for the cataloging and classification of one class of models, namely scientific models. Scientific models are defined as works and models classification is based on facet analysis.

1. Introduction

Scientific models can be used to improve science learning and integrate scientific research with teaching. NASA (1988) summarizes the kinds of prerequisites that models useful for science learning, specifically geo-sciences, must enable. These include:
1. Acquisition of observations – the model must exhibit a selective attitude to information;
2. Analysis and interpretation of the observational data – the model must be structured, pattern-seeking and replicating;
3. Construction of and experimentation with conceptual and numerical models; and
4. Verification of the models - reliability and validity for further testing, experimentation, and replication.

ThinkerTools (2001) at the University of California at Berkeley has developed Morton Modeler, a computer agent, who walks students through the process of designing and building good scientific models. Projects such as the Alexandria Digital Library are trying to include models for providing learning spaces in digital libraries (Coleman, 2001). A metadata standard with approximately 150 elements of description has been developed for describing computational models in the geo-sciences (Hill, 2001).

2. Background

The word model has many definitions. In ordinary, everyday usage it can be used to indicate a three-dimensional reproduction, usually on a smaller scale, a design or style of structure, a person or thing that is worthy of imitation, a person hired to pose for an artist or display clothes, etc. (OED Online, 2001). Model can also be used as an adjective in the sense of exemplary.

In the sciences and social sciences, the word model has different connotations and this study limits its analysis of models to the sciences. Minshull (1975) argues that a model in geography can be a theory, law, hypothesis, structured idea, a role, relation, equation, reasoning, or synthesis of data. Representational objects such as maps and satellite images are also considered models (Chorley and Haggett, 1967). The increasing use of information technology and the advent of the World Wide Web (WWW) has added another important
dimension to the scientific modeling enterprise. Many models are now available as black box models and web applets. Black box models conceal complex computational and other features of modeling while applets typically do not require special computers to run the models. Visualizations, simulations and datasets are also increasingly being referred to as models; sometimes these are products (outputs) and inputs of the activity of scientific modeling.

Models are the intellectual artifacts, the creative products of scientific research and important tools for learning science skills. (Heckhausen, 1972, 85). Yet libraries, bibliographic databases, and the Internet/WWW do not provide collocated and easy access to models. A partial list of problems in the information retrieval of models in one area, constitutive relations in Civil Engineering, includes:

1. Vocabularies used currently in the major databases for the discipline do not have or use a descriptor constitutive models.

2. There is a lack of terminological consistency across different sources. Information resources and packages such as books, theses, dissertations, and reports, about models are described in library catalogs using the controlled vocabulary scheme of Library of Congress Subject Headings (LCSH). Smaller information packages such as articles from journals and conference proceedings are covered by indexing and abstracting sources and bibliographic databases. For these, the preferred controlled vocabulary varies from database to database. For example, COMPENDEX is one of the major databases in engineering and it uses its own scheme. Therefore, users must become familiar with at least two different vocabularies to search for constitutive models in the library catalog and a bibliographic database or periodical index.

3. Unmatched knowledge structures: Vocabularies map the knowledge of experts in a discipline; this is usually the authors and writers of the articles indexed and based on literary warrant. The knowledge structure thus derived is often significantly different from practitioners and learners who may have incomplete knowledge structures. An informal survey of Civil Engineering faculty at the University of Arizona at Tucson showed that for improving learning constitutive models as one cohesive class under which items with different facets/aspects about constitutive modeling could be further sub-arranged or displayed was preferred. For example, rather than specific subjects under constitutive models such as creep, stress and strain, faculty would have preferred model facets/aspects such as a typology (elastic, plasticity-based, discontinuity, novel) and object modeled (tunnel, slope, etc.).


Thus, the overarching question can be stated quite simply, what is a scientific model and how can it be represented in the library catalog to reveal disciplinary structures and facilitate information retrieval for learning?

3. Research Questions

This models classification project is part of a larger study that is developing a
registry and prototype database/catalog of scientific models. Findings for the research questions that are explored in this paper are:

1. Are scientific models works?
2. What are the facets in a scientific models classification scheme?

Other subsequent research questions that the study will be exploring include:

1. What is a models classification scheme?
2. What are characteristic descriptive properties of models that can be automatically derived as a function of model classification?
3. How can such a classification/nomenclature be used in 1) retrieving information about models, 2) archiving, and 3) model execution via the WWW browser.

4. Hypothesis

It is hypothesized that a scientific model is a work. It is further hypothesized that classification or nomenclature for models can serve best as the basis of models description and retrieval in digital libraries with learning spaces. Classification is making a re-emergence as a viable method for the semi-automatic organization of electronic resources. See for example, Project DESIRE (Koch, 1997) and Project Scorpion (The Scorpion project, 2002). Because scientific models are complex objects that have multiple representations, exist in heterogeneous forms and formats, and include interdisciplinary problems and domains, classification may be useful in improving retrieval of models by revealing implicit knowledge structures, subject and other relationships.

5. Methods

A preliminary study, to test the notion of scientific models as works, was conducted. It included document, content analysis, as well as retrieval of selected scientific models in a bibliographic utility, OPAC, controlled vocabulary and classification schemes, and bibliographic databases. Methods and preliminary results are fully documented elsewhere and readers are encouraged to refer to it or contact author for details (Coleman, in press).

6. Results

In the interest of brevity only results from two specific parts of the study are presented here. First are the findings from the document analysis regarding the physical composition of one scientific model, Atmosphere-Ocean Model from NASA at the URL: http://aom.giss.nasa.gov/. Scientific models appear to have the following forms and formats:

1. Theory or Hypothesis (text file or graphic diagram/image)
2. Observations and Measurements (data sets in various formats)
3. Computer Hardware (descriptions only)
4. Computer Software (this includes actual software executable code, other software essential for running the model code itself, documentation (technical and user guide that accompanies the software)
5. Animations (images, video)
6. Visualizations (images, video)
8. Reports, Reviews, Annotations, Experiments, Articles (text)
9. Tools (applets, services)
10. Mathematic (algorithmic, numeric notation)
11. FAQs, Mailing Lists/Listservs, Discussion Groups, Events (various).

Second, a summary of the findings of the content analysis of published literature (texts) shows that the following facets are present in models in the area of water quality:
1. Concept: is an idea, the traditional subject (for example, calculus of variations)
2. Object: the object studied in the model
3. Discipline: the major discipline to which this model belongs (may be determined either through author affiliations or other means)
4. Phenomenon: the phenomenon being modeled
5. Process: the process being modeled
6. MathRepresentation: the mathematical functions, equations used
7. Software: the software needed to run the model
8. FundamentalLaw: the fundamental laws that the model is based upon
9. Type: the type of model based on its purpose
10. Variable: number, types, conditions, and variables in this model
11. Problem: the problem the model is analyzing stated often as a question
12. Theory: is there an existing theory or research group to which this model belongs

7. Discussion

Drawing from information management traditions in libraries, archives, and museums, there are at least three perspectives that can be used for organizing scientific models as works. The library tradition emphasizes collection of individual objects, the independent bibliographic unit, or the individually available information package. If this framework were to be used, the emphasis would be on descriptive cataloging. The archival tradition emphasizes collections of related objects. The museum tradition emphasizes consistent, constant, change. Disparate objects are often collected as exhibits based on changing or newly revealed and articulated relationships.

Examining scientific models as works provides a strategy that appears to draw the best from all three traditions noted above. It is described as a series of practical steps in the development of a classification scheme for models.

Step 1: “A work is the intellectual content of a bibliographic entity; any work has two properties: a) the propositions expressed, which form ideational content and b) the expressions of those propositions (usually a particular set of linguistic (musical, etc.) strings) which form semantic content.” (Smiraglia, 2001). Semantic content in scientific models includes mathematical expressions, formal propositions and hypotheses, and statements of laws. Ideational content includes ideas about objects, processes, and relationships, usually within or for specified spatial and temporal scales, and formally, semantically expressed as mathematical equations and algorithmic notation. The ideas include both observables (verified and expressed as measurements) and non-observables (hypothetical data, mathematical equations).

Step 2: A work is a bibliographic entity. Scientific models as bibliographic
entities have two properties: physical and conceptual.

Step 3: The physical components of a scientific model can be determined in terms of its form (what the instantiation is) and the following forms and formats of models exist

1. Textual works – includes articles, abstracts, bibliographies, reviews, analysis, software documentation.

2. Datasets – includes observations and measurements of the observed phenomenon, object, process reported as data, images, visualizations, and graphs.

3. Software – includes computer code, both source code and downloadable executables.

4. Services – includes interactive and other services (animation applets, databases, indexes, contact pages, submit forms, etc.)

When conceptual components are examined the facets, the basic ideas, the model expresses, can be abstracted. Even more than just the ideas, the ideational (subject + other) relationships are important in modeling. Conceptual components are called model concepts and relationships and the 12 facets from the results section (see section 6 above) are the preliminary facets for subsequent development.

Step 4: We are developing a prototype catalog of scientific models using this preliminary classification. We are using the Dublin Core Metadata Element Set, Version 1.1, as the base metadata scheme for cataloging selected models. In addition, the 12 models facets are being framed for the area of water quality with other general additional facets such as form, time, space, author/group and bibliographic relations. A prototype database should be available by the end of the year 2002. A sample framework is shown below.

<table>
<thead>
<tr>
<th>A Models</th>
<th>rapids</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Scientific models</td>
<td>5 (by mathematical representation)</td>
</tr>
<tr>
<td>2 water</td>
<td>51 differential equations</td>
</tr>
<tr>
<td>21 rivers</td>
<td>6 (by software)</td>
</tr>
<tr>
<td>22 lakes</td>
<td>use MIME formats here</td>
</tr>
<tr>
<td>3 (by process)</td>
<td>7 (by fundamental law)</td>
</tr>
<tr>
<td>infiltration</td>
<td>8 (by discipline)</td>
</tr>
<tr>
<td>4 (by phenomenon)</td>
<td>9 (by concept) – use traditional subject headings here</td>
</tr>
</tbody>
</table>

Step 5: Each unit of analysis is given a class number for each facet in addition to the metadata that is created for the unit. Most of the facets are currently using controlled values from different schemes until the classification framework is completed.

8. Conclusion

Scientific modeling conceals an incredible amount of intellectual relationships that traditional bibliographical tools, primarily the catalog and the index, neither capture nor describe from the texts, documents, and items about models. Our increasing awareness of conceptual and textual instability of electronic forms requires active investigation and experimentation with other types of knowledge.
organization and representation structures. This paper has presented an alternative solution to view scientific models as *works* and use a models faceted classification scheme for their subsequent display and information retrieval. Decades of research both in information retrieval and information seeking behavior complemented by the widespread success of Internet search engines has shown us that users tend to disregard Boolean searches, human indexing as opposed to machine indexing does not improve search performance significantly, and that users want a few relevant, good materials. Assessing relevance in terms of disciplinary structures has never been researched. Therefore the subsequent use study of the developed models catalog/database and the evaluation of retrieval and displays based on classification should reveal interesting findings.

Acknowledgements

The comments of three anonymous reviewers who approved the funding for the models classification study is gratefully acknowledged. Feedback from the anonymous reviewers of the Seventh International ISKO Conference review panel were also very useful and I thank them. I thank my former student, Olha Buchel, for many meaningful discussions on the cataloging and classification of models and for her support in classification, database, and catalog design. Thanks to University of Arizona Professor Muniram Budhu, Dept. of Civil Engineering, Dr. William Rasmussen, Dept. of Agriculture and Hydrology, and Dr. John Kemeny, Dept. of Rock and Mining Engineering for their participation in the informal survey of models for teaching and learning engineering concepts. This study is supported by the University of Arizona, Office of the Vice President for Research Faculty Small Grant, 2001-2002

References


Thinkertools; *The Thinkertools Scientific Inquiry and Modeling Project*, University of California at Berkeley. [http://thinkertools.soe.berkely.edu/]. Accessed 01/31/02.
Indexing and Querying of Narrative Documents, a Knowledge Representation Approach

Abstract: We describe here NKRL (Narrative Knowledge Representation Language), a semantic formalism for taking into account the characteristics of narrative multimedia documents. In these documents, the information content consists in the description of 'events' that relate the real or intended behaviour of some 'actors' (characters, personages, etc.). Narrative documents of an economic interest correspond to news stories, corporate documents, normative and legal texts, intelligence messages, representation of patient's medical records, etc. NKRL is characterised by the use of several knowledge representation principles and several high-level inference tools.

1. Introduction

Narrative documents, or 'narratives', are multimedia documents that describe the actual (or intended) state or behaviour of some 'actors' (or 'characters', 'personages' etc.). These try to attain a specific result, experience particular situations, manipulate some (concrete or abstract) materials, communicate with other people, send or receive messages, buy, sell, etc.

Leaving fiction documents aside, we can note that:

- A considerable amount of the natural language information that is relevant from an economic point of view deals with narratives. This is true for all the different sorts of news story documents, for most of corporate information, for the intelligence messages, medical records, notarised deeds, legal documents, etc.
- In narrative documents, the actors or personages are not necessarily human beings. We can have narrative documents concerning, e.g., the vicissitudes in the journey of a nuclear submarine (the 'actor', 'subject' or 'character') or the various avatars in the life of a commercial product.
- It is not even necessary that the narrative situations to deal with be recorded in natural language (NL) documents: they can also be represented, e.g., as Web images, video or digital audio documents.

We introduce here NKRL, the "Narrative Knowledge Representation Language" (Zarri, 1997; 1998), a language expressly designed for representing the conceptual gist of economically relevant narratives. NKRL has been used as 'the' modelling language for narratives in several European projects; see, e.g., the recent Concerto (Esprit P29159) and Euforbia (IAP P26505) projects.

2. The main knowledge representation tools used in NKRL

Each 'component' of the set of knowledge engineering representation tools of NKRL takes into account a specific aspect of the global narrative domain.

The "definitional component" supplies tools for representing the 'concepts', intended here as a formal representation of the 'important notions' of a given application domain. A concept is rendered, substantially, as a frame-like data
structure associated with a symbolic label like \textit{human\_being}, \textit{taxi\_}, \textit{city\_}, \textit{chair\_}, \textit{gold\_}, etc. Concepts are inserted into a generalisation/specialisation hierarchy that is called \textit{H\_CLASS(es)}, and which corresponds well to the usual ontologies of terms; see (Zarri, 1997) for additional details.

The "enumerative component" of NKRL concerns the formal representation, as instantiated frames, of the concrete realisations (\textit{lucy\_\_}, \textit{taxi\_53\_}, \textit{paris\_\_}) of the concepts inserted in the \textit{H\_CLASS} ontology. In NKRL, these instances take the name of \textit{individuals}. Throughout this paper, we will use the italic type style to represent a \textit{concept\_}, the roman style to represent an \textit{individual\_}.

The "descriptive component" concerns the tools used to produce the formal representations (called "templates") of general classes of narrative events, like "moving a generic object", "formulate a need", "be present somewhere", etc. In contrast to the traditional \textit{ternary} (name-attribute-value) frame-like structures used for concepts and individuals, see \textit{H\_CLASS}, templates are characterised by a \textit{quaternary} format connecting together, essentially, the symbolic name of the template, a predicate and several arguments of the predicate. These last are, in turn, differentiated through the use of a set of named relations, the roles. If we denote with \textit{L\_} the generic symbolic label identifying a given template, with \textit{Py} the predicate, with \textit{R\_*} the generic role and with \textit{a\_*} the corresponding argument, the template data structures have then the following format:

\[
(L_t (P_j (R_1 a_1) (R_2 a_2) \ldots (R_n a_n))).
\]

The instances (called "predicative occurrences") of the templates, i.e., the representation of specific elementary events like "Tomorrow, I will move the wardrobe" or "Lucy was looking for a taxi" are in the domain of the last component, the factual one.

3. Using templates and occurrences to represent concrete events

To represent a narrative like "British Telecom will offer its customers a pay-as-you-go (payg) Internet service in autumn 1998", we must select firstly the template (descriptive component) corresponding to "supply a service to someone", see the upper part of Table 1. This template is a specialisation of the particular \textit{move} template of \textit{EMTEMP} corresponding to "transfer of resources to someone". In a template, the arguments of the predicate (the \textit{a\_*} terms in (1)) are represented by variables with associated constraints — that are expressed as concepts or combinations of concepts, i.e., using the terms of the \textit{H\_CLASS} hierarchy. The constituents (as \textit{SOURCE} in Table 1) included in square brackets are optional.

When deriving a predicative occurrence, like \textit{c1} in Table 1, the role fillers in this occurrence must conform to the constraints of the father-template. For example, in occurrence \textit{c1\_}, \textit{british\_telecom} is an individual instance of the concept \textit{company\_} that is, in turn, a specialisation of \textit{human\_being\_or\_social\_body\_}, etc. The meaning of the expression "\textit{BENF (SPECIF customer\_\_british\_telecom)}" in \textit{c1} is self-evident: the beneficiaries (role \textit{BENF}) of the service are the customers of — \textit{SPECIF (ication)} — British Telecom.

The (about 200) templates that make up actually the \textit{H\_TEMP} hierarchy —
the ‘catalogue’ of NKRL templates — are permanent and fully defined. We can say that these templates are part and parcel of the definition of the language; an (extremely simplified) image of H_TEMP is given in Figure 1. This approach is particularly advantageous for practical applications because it implies that: i) a system-builder does not have to create himself the structures needed to describe the events proper to a (sufficiently) large class of narratives; ii) it becomes easier to secure the reproduction or the sharing of previous results.

name: Move:TransferOfService
father: Move:TransferOfResource
position: 4.231
NL description: 'Transfer or Supply a Service to Someone'

OWN        SUBJ         var1: [var2]
OBJ         var3
[SOURCE     var4: [var5]]
BENF        var6: [var7]
[MODAL      var8]
[TOPIC      var9]
[CONTEXT    var10]
{[modulators]}

var1 = <human_being_or_social_body>
var3 = <service_>
var4 = <human_being_or_social_body>
var6 = <human_being_or_social_body>
var8 = <action_name>
var9 = <sortal_concept>
var10 = <event_>
var2, var5, var7 = <physical_location>

Table 1. Deriving a predicative occurrence from a template.

The basic NKRL tools are enhanced by the use of two additional mechanisms:
• the AECS ‘sub-language’ that allows the construction of complex (structured) predicate arguments called ‘expansions’;
• the second order tools (binding structures and completive construction) used to code the ‘connectivity phenomena’ (logico-semantic links) that, in a narrative situation, can exist between single narrative fragments.

Because of the space restrictions, it is now impossible to supply detailed information about these two mechanisms, see (Zarri, 1997; 1998) for further details.
4. Some remarks on the queries and inference procedures

Search patterns are NKRL data that supply the general framework of information to be searched for, by filtering or unification, within an NKRL knowledge base. The upper part of Table 2 is the representation of a very simple narrative fragment: “On June 12, 1997, John and Peter were admitted \( (\text{together} = \text{COORD(ination)}) \) to hospital”. The “temporal modulator” \text{begin} asserts that the date associated with \text{date-1} corresponds to the beginning of the state of being at the hospital. Modulators (deontic, modal, and temporal modulators) are special codes that are added to the basic core of a predicative occurrence to better specify its conceptual meaning, see (Zarri, 1998).

A simple example of search pattern, translating the query: “Was John at the hospital in July/August 2001?” is represented in the lower part of Table 2. The two timestamps associated with the pattern constitute the “search interval” that is used to limit the search for unification to the slice of time that it is considered appropriate to explore.

In the Java XML/RDF-compatible version of NKRL (see Zarri, 2000), the FUM (Filtering Unification Module) deals with the unification of search patterns. Its inference level is only a first step towards more complex reasoning strategies, see (Zarri and Azzam, 1997). A powerful class of NKRL inference rules concerns the ‘hypotheses’, see the example of Table 3. We suppose here to have directly retrieved, thanks to FUM, a given information within a base of NKRL occurrences, e.g., the information: “Pharmacopeia, an USA biotechnology company, has received 64,000,000 dollars from the German company Schering in connection with an R&D activity”.

We suppose, moreover, that this occurrence is not already explicitly related with other occurrences of the base. Under these conditions, we can activate a specific Java module, Inference Engine that, using a rule like that of Table 3, will try to rely automatically the information found originally by FUM with other information in the base. If this is possible, this last information will represent a sort of ‘causal explanation’ of the information originally retrieved — in our example, an ‘explanation’ of the money paid by Schering. We will find, e.g., that “Pharmacopeia and Schering have signed an agreement concerning the production by Pharmacopeia of a new compound” and that “In the framework of the agreement previously mentioned, Pharmacopeia has actually produced the new compound”. 
Figure 2. A (very reduced) version of the H_TEMP Hierarchies

c2) EXIST SUBJ (COORD john_peter_): (hospital_1)
{ begin }
date-1: 2-june-1997
date-2:

(?w IS-PRED-OCCURRENCE
:predicate EXIST
:SUBJ john_
:location of SUBJ hospital_
(1-july-1997, 31-august-1997))

Table 2. An example of search pattern.
5. Conclusion

In a ‘traditional’ ontology, concepts are defined (synthetically) as frames. It is now evident that an organization in terms of frames is largely sufficient to provide a static definition of the concepts — i.e., a definition \textit{a priori} of each concept considered in itself. It is, however, difficult to admit that this type of organization could be sufficient to define the dynamic behaviour of the concepts, i.e., to describe their mutual relationships when they take part in some concrete action, situation, etc. (the NKRL ‘events’).

\begin{center}
\begin{tabular}{l}
\textbf{Premise:} \\
\textbf{RECEIVE} \quad \textbf{SUBJ} \quad x \\
\textbf{OBJ} \quad \textbf{money} \\
\textbf{SOURCE} \quad \textbf{y} \\
\end{tabular}
\end{center}
\begin{center}
x = \textit{company} \\
y = \textit{human\_being} \mid \textit{company} \\
“A company has received some money from another company or a physical person”
\end{center}

\begin{center}
\begin{tabular}{l}
\textbf{First condition schema:} \\
\textbf{PRODUCE} \quad \textbf{SUBJ} \quad (\textit{COORD x y}) \\
\textbf{OBJ} \quad z \\
\textbf{BENF} \quad (\textit{COORD x y}) \\
\textbf{TOPIC} \quad (\textit{SPECIF process \_ v}) \\
\end{tabular}
\end{center}
\begin{center}
z = \textit{mutual\_commitment} \mid \textit{business\_agreement} \\
v = \textit{tool\_product} \\
“A general or business-oriented agreement about the creation of a new product has been concluded by the two parties mentioned in the premise”
\end{center}

\begin{center}
\begin{tabular}{l}
\textbf{Second condition schema:} \\
\textbf{PRODUCE} \quad \textbf{SUBJ} \quad x \\
\textbf{OBJ} \quad \textbf{v} \\
\textbf{MODAL} \quad w \\
\textbf{CONTEXT} \quad z \\
\end{tabular}
\end{center}
\begin{center}
w = \textit{industrial\_process} \mid \textit{technological\_process} \\
“The company that received the money has actually created the product mentioned in the first condition schema”
\end{center}

\begin{table}[h]
\centering
\begin{tabular}{|l|}
\hline
\textbf{Table 3. An example of high-level inference rule in NKRL.} \\
\hline
\end{tabular}
\end{table}

Reducing the description of \textit{events} to the description of \textit{concepts} — as has been sometimes proposed — is nothing more than a further manifestation of that well known ‘uniqueness syndrome’ proper to some Artificial Intelligence and Knowledge Representation milieus. In NKRL, we make use in an integrated way of several sorts of representational principles (more details in the final and complete
paper) and several years of successful experimentation with the most different
narrative situations are there to testify that this seems not to be a totally
unreasonable approach.

References
Zarri, G.P. (1997) NKRL, a Knowledge Representation Tool for Encoding the 'Meaning' of
Complex Narrative Texts, Natural Language Engineering — Special Issue on
Knowledge Representation for Natural Language Processing in Implemented Systems,
3, 231-253.
Its Potential for Legal Narratives, Information & Communications Technology Law —
Special Issue on Models of Time, Action, and Situations, 7, 213-241.
Business-Oriented Domains", in Industrial Knowledge Management: A Micro-level
Zarri, G.P., and Azzam, S. (1997) "Building up and Making Use of Corporate Knowledge
Repositories", in Knowledge Acquisition, Modeling and Management - Proceedings of
Interdisciplinary Knowledge Integration and Intellectual Creativity

Abstract: The use of culturally prominent metaphors, symbols, archetypes, myths, and narrative patterns as metadata is explored and analyzed as a method to facilitate the discovery and retrieval of information and the integration of knowledge across both disciplinary and cultural boundaries in order to promote intellectual creativity and interdisciplinary innovation. The rationale for metaphorical and symbolic metadata is to be found in recognition of the role of metaphorical and analogical thinking in intellectual creativity as well as in limitations of classificatory and disciplinary subject languages and the ontologies on which they rest. A Universal Cultural Symbol Thesaurus is described as a potential enumerated subject language for a usable lexicon of metaphors and symbols that have cognitive connotations as well as cultural and psychological resonance. Such a thesaurus could be employed to classify and index information objects in a symbolic dimension that would complement and run transverse to existing analytic and disciplinary modes of classification and fit easily into Boolean search logic.

1. Introduction

That the volume of stored and networked digital information is not only growing exponentially but being made interoperable - not only within the library world through such developments as Z39.50 but in the entire world through the spread of XML and the development of the Semantic Web (Berners-Lee, Hendler, & Lassila, 2001) -- poses a twofold and contradictory challenge to knowledge integration. On the one hand, clarity, coherence, and universality are crucial to the organization and representation of knowledge in order to make it maximally accessible and discoverable in useful ways. On the other hand, since all knowledge organization reflects ontology, more universal means of organization may squeeze all the knowledge in the world into a hegemonic ontology that is at odds with the creative, improvisatory, and serendipitous processes through which new knowledge is created and possibly with the ontological assumptions and commitments of local cultures, despite the Semantic Web project's explicit commitment to preserving the ontologies of particular communities.

We know that intellectual creativity often occurs through processes of analogical and metaphorical thinking that step over the neat boundaries of official and inherited disciplines and classification schemes. Certainly, in the humanities and social sciences, new knowledge arises as much from the bricoleur, who patches things together out of often dissimilar materials that are at hand, as from the instrumentally rational "engineer" of knowledge. In what follows, I address, through a particular take on universality, the ISKO conference objective of proposing new models, methods, and techniques of integrating knowledge across both disciplinary and cultural boundaries. My focus is on creating a metadata set of universal or at least widespread and well-known cultural metaphors and symbols,
whose use would amount to a new method of knowledge organization and retrieval and might contribute to interdisciplinary knowledge creation in the social sciences and humanities. In particular, a Universal Cultural Symbol Thesaurus is described as a potential subject language: a lexicon of metaphors and symbols that have cognitive connotations as well as cultural and psychological resonance and that would enhance bricolage and serendipity as paths through the gigantic storehouse of digital information.

2. Theoretical background

An important body of theory and research in psychology, cognitive science, sociology, anthropology, and philosophy, from a variety of conceptual frameworks and points of departure, identifies a duality - or more accurately a set of related and analogous dualities -- in human cognition, culture, and social organization that is of immense significance for the theory and practice of classification: the duality that Goody calls the "Grand Dichotomy" (Goody, 1977). This encompasses such dichotomies as rational/mythical, instrumental/expressive, scientific/narrative (Bruner, 1986), system/lifeworld (Habermas, 1988), disembodied/embodied (Dreyfus, 1992), engineer/bricoleur (Lévi-Strauss, 1966), high-focus/low-focus (Gelernter, 1997), allegory/analogy (Stafford, 1999), and related conceptual pairings, including of course modern/premodern (and modern/postmodern). In general, modernity has involved privileging or emphasizing the first member in each of these dichotomies and defined rationality and progress in its terms, de-emphasizing, suppressing, or eliminating the second.

The Grand Dichotomy is relevant to the organization of knowledge and information and to classification for two reasons: first, because it bears on the underlying ways in which knowledge and the world are structured and how they should be represented; second, because it bears on how people seek, acquire, and use information and knowledge. Every scheme for classifying or ordering information is grounded in or implies a philosophical ontology or cosmology. Overall, modern information organization is based on the first term in the above pairs. It is grounded in an "ideology" (Svenonius, 2000) based upon general systems theory, positivistically oriented philosophy of science, and linguistic philosophy that privileges instrumental and scientific rationality. Its dominant classification schemes (e.g. the LCSH and DDC systems) inherit from the Aristotelian Tree of Porphyry and Neo-Platonic, realist ontology and theology, the "Great Chain of Being". Especially in the Neo-Platonic version, this ontology asserts the priority of the universal over the particular and of the abstract over the concrete and sees the individual or particular as merely an emanation or instantiation of the abstract and the universal. Furthermore, the knowledge or information seeker has tended to be construed as an instrumentally rational, "disembodied" cognitive being pursuing clear, known information goals. This is reflected in conventional library profession definitions of information literacy.

Certainly, this model of knowledge organization and of the knowledge seeker corresponds to a genuine and important dimension of cultural reality and social, scientific, and individual needs. However, postmodern awareness of the limitations of objectivistic and rationalistic frameworks for representing and structuring knowledge has delegitimized the philosophical systems and assumptions that underlie it and points to the unstable and socially constructed nature of modern
information organization and classification schemes (Bowker & Star, 1999). Furthermore, the progress of the sciences leads to taxonomic complexification that is in principle unlimited (Rescher, 1998), which diminishes the utility and relevance of both linear and hierarchical models of classification and gives rise to new modes of cognitive inter-relationship and ordering. The new structure "is not that of a hierarchy at all, but rather that of chain-mail-work interlinkage reminiscent of medieval armor" (Rescher, 1979). This trend is amplified by the sheer volume of accessible digital information and by the emergence of hypertext as a novel and characteristically postmodern method of information ordering that, through the World Wide Web, has become a global system for organizing information and knowledge with a simple and viable, although associative, technical infrastructure. Through its use in personal publishing on the Web, hypertext has taken on tremendous cultural and psychological force for individuals, organizations, social groups, and information producers and managers. In its technical structure, it has given rise to two principles of organizing information: through properties of the text itself and its connection to other texts, and through properties of the individual who makes the links, either through construction and publication of the text or through its reception and deconstruction. Hypertext mirrors postmodernism's thesis of the personally and socially constructed nature of reality.

A core feature of the postmodern intellectual situation is a re-evaluation and critique of the dominance of the first term in the above dichotomous pairs and an attempt to rehabilitate or recognize the legitimacy of the second member or restore it to its rightful place: thus Habermas's attempt to undo the colonization of the lifeworld by the system and restore communicative rationality, Dreyfus's and others' attempt to restore embodiment to epistemology and to our relation to technology, Bruner's attempt to rehabilitate and legitimate narrative knowing, Stafford's rehabilitation of analogical thinking, Gelernter's design of computer systems capable of "low-focus", metaphorical, analogical functioning, and so on. In general, these postmodern approaches are based not on reversing the hierarchical ordering of the dichotomy but finding an appropriate, balanced relationship between what must be seen as two vital dimensions of individual, cultural, and social existence.

In this context, recent research emphasizes the metaphorical and analogical substructure of rational/conceptual thought, whether in the form of the evolutionary survival in the brain of the cognitive analog of the mythic phase of cultural evolution (Donald, 2001); Lakoff and Johnson's work on embodied thought and the metaphorical infrastructure of thought (Lakoff & Johnson, 1999); Fauconnier's work on conceptual blending and the role of metaphor in mapping between mental spaces; Durand's "transcendental fantastic" and general archetypology (Durand, 1992); or Stafford's argument that "by recuperating the sophisticated workings of ancient analogy for modern science, I believe artists and art, architecture, design, film, and media historians, in particular, can contribute a cross-cortical model of the complex processes of mental combination" (Stafford, 1999, 144) Thinking metaphorically and analogically is fundamental to creative thinking and intellectual innovation (Root-Bernstein & Root-Bernstein, 1999). Creativity and knowledge integration also involve serendipity and receptivity as well as the directed pursuit of specific cognitive goals. In the humanities and social sciences intellectual innovation often occurs through the synthesis or mutual amplification of ideas and information from disparate domains, sometimes linked associatively. There is reason to believe that this is true to some extent in the natural sciences as well.
Metaphors and symbols are often key vehicles for analogical thinking, the bridge of resemblance among diverse areas and phenomena.

3. Metaphorical and Symbolic Metadata

In postmodern, complex, multi-cultural society, any semblance of a universal, background cosmology, cultural system, or generally shared lifeworld that could serve as an accepted common basis for structuring knowledge and information has dissolved, leaving in its place a multiplicity of diverse lifeworlds, orientations, and individual meaning schemes (Habermas, 1992). Thus, it would be futile to try to invent a new, more encompassing and universal classification system or taxonomy that would hierarchically subsume all conceptual frameworks. Metaphors and cultural symbols, on the other hand, provide ways of linking across lifeworlds and cultural systems while expanding intercultural understanding, especially if they are drawn from a variety of cultural traditions.

What is metaphorical or symbolic metadata? In brief, they are conventional signs or symbols attached to data in order to ascribe metaphorical or symbolic meaning to it. In the present context we limit ourselves to signs or symbols drawn from culturally significant symbol systems or sets, such as the I Ching, the Tarot, astrology, the Kabbalah, Yoruba mythology, Christian iconography, etc. that have known traditional meanings. Thus, Kant's Critique of Pure Reason could be described by the sun, a symbol of light and enlightenment; a biography of Gustav Mahler could be described by a red rose, a symbol of martyrdom. There are really several kinds of symbols that are candidates for use as metadata, some of which overlap: symbols (e.g. the sun or a rose), concepts (as in the I Ching hexagram for "the creative"), archetypes (e.g. the Virgin), myths (e.g. the myth of Sisyphus), narrative patterns (typical story structures), and narrative or action characters (e.g. the journeying hero). In general, symbolic metadata would classify the knowledge element or information object (book, article, Web page, art work, etc.) as belonging to the semantic field connotated by the symbol or metaphor in question.

The aspect of a metaphorically or symbolically based classification system that is most difficult to swallow is the inevitable element of arbitrariness and subjective interpretation involved in assigning a symbolic descriptor to an information object or bibliographic record. While ambiguity, interpretation, and subjectivity are involved to some extent in classification generally, a principal merit of standard classification systems is the existence of conceptually and analytically specifiable criteria for the assignment of terms. Metaphors and symbols do have identifiable and specifiable meanings, recorded in symbol dictionaries such as (Ferguson, 1954), that could serve as a model for a symbol thesaurus; otherwise, they would not preserve their meanings and identities. However, by their very nature and use their boundaries are fuzzy or porous. It is precisely this that enables them to be applied to new things and to be used creatively and in analogical thinking.

The tension between metaphorical terms and the analytical descriptors of conventional subject languages derives from underlying ambiguities and difficulties in the fundamental notion of "aboutness", as expounded so well by Svenonius (2000). She differentiates between the grammatical model of aboutness, oriented toward subject analysis as a process of summarization and based on the logical positivist idea of propositions as providing a picture of what there is, and a model
that recognizes that language is often used non-propositionally and non-referentially and that subjects do not always have names. In Svenonius's words, "the scientific model of aboutness is limited, as indeed is subject analysis itself" (2000, p. 48). Of course, the fuzziness of metaphors and symbols is an impediment to automating subject determination, since every act of classification is an act of interpretation in the emphatic sense, and the tremendous social, cultural, and technological pressure to automate will be a potent factor in limiting receptivity to or enthusiasm for the idea of a metaphorical subject language. In the limiting case in which the sentences in a document provide no basis for determining what metaphors or symbols might be used to describe it, it is currently impossible to envision automating subject analysis. Nevertheless, if we follow Svenonius in taking a painting or a piece of music as an information object about which the subject is difficult to determine, even with such works there is often critical agreement about the range of meanings that can be attributed to them. Non-existent or rare are those music critics who would describe Beethoven's Ninth Symphony as "cute" or his Sixth Symphony as "morose". Non-propositional meanings and symbols may be fuzzy and ambiguous, but they are not empty or arbitrary.

How would one use metaphorical and symbolic metadata? For bibliographic and information classification and retrieval they would rarely be used on their own, but rather transversely or orthogonally to standard classification schemes. For example, a scholar venturing into disciplines other than her own might search for information that was indexed under a symbol of personal, cultural, or intellectual interest, or might use such a symbol, through a Boolean AND, to select data, on analogical grounds, out of a large set of information objects: locate a journal article with author x AND keyword y AND symbol z. An individual who had identified her own creativity as particularly focused on or stimulated by a particular symbolic theme might use symbolic or metaphorical metadata to find material to nourish her creativity or imagination. Communities attempting to formulate their concerns and influence the political process might use symbolic or metaphorical metadata to search for relevant information. Students trying to articulate or formulate the motivations guiding their studies might find that congenial symbols serve as a path of entry into a mass of unfamiliar material. It is worth noting that symbolic metadata could be assigned for both cognitive/denotative reasons, i.e. to denote the semantic content of the information object, and expressive/connotative ones, for example to characterize the subjective meaning of the knowledge element to its creator. This could be handled by assigning two symbolic metadata fields to a bibliographic record, one for each, thus enabling individuals to search for information objects based on their subjective meaning to their creators as well as on meanings determined by professional librarians or classifiers. Of course, as in standard classification systems, a number of symbolic descriptors could be ascribed to a single information object.

4. A Universal Cultural Symbol Thesaurus

How could metaphorical and symbolic metadata be implemented in a usable manner? A model suggested by the tradition of both folk and scholarly symbol dictionaries is a Universal Cultural Symbol Thesaurus (UCST). This would draw on the symbol traditions of diverse cultures representing different geographical, ethnic, language, and religious groups. Like existing symbol dictionaries, the UCST would,
for each symbol, provide a number, an image, and a short description of the meaning - or multiple meanings - of the symbol, which would need to be determined by folklorists, anthropologists, and scholars of religious and cultural traditions. As discussed above, such meanings cannot be exhaustively specified or definitively bounded, since they are always mediated by hermeneutic expansion and subjective interpretation; nevertheless, recurrent and semi-stable meaning patterns have been identified. Ideally, the UCST would need to contain enough symbols to capture major elements from multiple traditions and yet be short enough so that individuals could become familiar with it in its totality, for purposes of either classification or retrieval: 256, a good, round computer number, suggests itself for an initial implementation. It could take as its point of departure a universal "archetypology", such as Durand's "anthropology of the imagination" (1992), which already draws on a vast cross-cultural literature, and amplify it with other cultural materials under the guidance of anthropologists and folklorists. To test the potential value of such a thesaurus, it should be used to classify a corpus of information objects. Usage and retrieval patterns could then be studied.

Since creative thinking, intellectual power, and theoretical innovation in the humanities and social sciences often occur through trans-domain, cross-literature, or interdisciplinary leaps or through serendipitous discovery based on analogy or parallelism, the addition of metaphorical or symbolic metadata to standard classificatory schemes could facilitate such leaps and discovery. For a researcher would be able to search for and retrieve information based not only on categories derived from familiar domain ontologies, thesauri, and classification schemes but also on shared symbolic or metaphorical meanings, both cultural and personal.

Acknowledgments
I am grateful to Shelley K. Hughes, Lydia Foerster, Linda F. Crafts, and Michael F. McCullough for discussing with me the issues raised in this paper and helping me clarify my thinking, and to Bob Silverman for his constant insistence on the "in-between."although the profession seems to be moving toward a more open-ended definition as "the ability to locate, evaluate, and use information to become independent life-long learners" (Kirk, 2001).

References


Rahmatollah Fattahi and Mehri Parirokh
Department of Library and Information Studies, Ferdowsi University of
Mashhad, Mashhad, Iran

Restructuring the Bibliographic Record for Better
Organization, Management, and Representation of
Knowledge in the Global Online Environment: A New
Approach

Abstract: The computer technology has considerable potential for better and more efficient
management of knowledge and information sources. It is possible to provide easy, integrated and straightforward access to works of knowledge appearing in different expressions, editions, and manifestations (i.e., to bibliographic families). The online environment requires that the bibliographic record, as the surrogate for works of knowledge, should be restructured to provide users with a high-quality finding, identifying and collocating bibliographic tool. Unlike the existing flat structure, the multi-level structure of the Work Record would provide catalog users with the ability to discover if a work has been published/represented in different editions and manifestations and if there is any relationship among different works in a catalog or on the Internet. A similar structure for the Author Record would provide better access to different works by an author or about an author. A new structure for the bibliographic record is suggested through an analysis and categorization of the different relationships that exist in the bibliographic universe. A prototype catalogue developed based on the new structure (available on the Internet at: http://wilma.silas.unsw.edu.au/ students/rfattahi/super.html) illustrates the ability of providing multi-level, yet easy access to voluminous works and voluminous authors. Furthermore, the prototype catalog provides a clearer picture of bibliographic families and helps OPAC users decide on specific editions, manifestations, or items related to a specific work.

1. Introduction
The value of information and knowledge in research and development (R&D) has extensively been discussed in the literature. The major source of accessing information and knowledge is, to a large extent, information retrieval (IR) systems mainly library catalogs and journal article databases. However, little has been written or researched about the ways in which IR systems can facilitate R&D through providing efficient access to knowledge rather than access to raw (bibliographic) information. This paper proposes that bibliographic information can be transformed into knowledge if the structure of the record and the catalog is redesigned to enable the system to fulfill users’ expectations in finding, integrating, identifying and understanding information as well as evaluating it in a given context. Current online library catalogs are far behind such functions.

2. Major problems with the existing online retrieval systems
For the last three decades that libraries have been using computers for information storage and retrieval, both end-users and librarians have faced
fundamental problems regarding retrieval relevance (i.e., too many hits, poor results, false drops, irrelevant documents, and poor displays) (Borgman, 1986, 1996; Wiberley, Daugherty and Danowski, 1995; Carlyle, 1999). This problem has exacerbated with the overwhelming explosion of data and information particularly in large catalogs and databases in that nothing is represented to the user as knowledge.

From the viewpoint of systems’ design, little progress has been made with current OPACs in recent times and most of the existing retrieval systems fail to rank, categorize and display the information retrieved on a well and intelligible basis. One of the most valid criticisms is that the pattern for organizing and displaying information makes little sense to users’ understanding. As Carlyle (1999: 3) points out, organized displays have seldom been incorporated into online catalog designs. The relationships between and/or among works are often obscured by the listing of irrelevant records among relevant ones or by the listing of related records with no information indicating relationships among the relevant records retrieved. Thus, the added value of information that could lead to the development of knowledge is lost in existing online systems.

Little research has been carried out to challenge the issue of knowledge organization and representation in online catalogs. Fattahi (1996a), Yee and Shatford Layne (1998), Carlyle (1999) are among the few researchers interested in such an area. IR systems need considerable research to identify the requirements for their development as knowledge systems. A major challenge of research in the area of information storage and retrieval has been, and still is, how to improve retrieval in terms of knowledge organization and representation.

3. Restructuring the bibliographic record: a new approach to make the online catalog a knowledge gateway

The need for research in restructuring the bibliographic record and the catalog has been stressed by some librarians, for example by Fattahi (1996a, 1996b), Tillett (1991, 1992), and Carlyle (1998). Information can be transformed into knowledge if the structure and pattern of the bibliographic record and the catalog is redesigned to enable the system fulfill users’ expectations in finding, identifying, relating and understanding information as well as evaluating it. As a surrogate for works, the bibliographic record is considered to be the main tool for fulfilling the different functions of the catalog. The set of data elements stored on a bibliographic record indicate physical and topical description as well as the nature of the work represented. These elements are recorded on a flat structure which is less relevant to functions such as identifying, finding, relating, organizing and sorting of information.

To provide catalog users with knowledge rather than unstructured information, the bibliographic record is in need of restructuring. Following are the proposed and tested ways in which the organization and displaying of all the instances in a bibliographic family can be managed in a fashion understandable to catalog users. The following approach proposes that, instead of displaying long lists of retrieved records which are hard for the user to browse, the contents of retrieved records can be categorized and represented in one or two screens as illustrated below. The Author Record and the Work Record are new approaches that help the
online catalog organize retrievals and achieve its functions.

4. The Author Record

The Author Record is a record which contains the title of the works created by or about an author. It can have a simple arrangement of works by the author (for example, alphabetical, chronological, or according to genres) that are available in any physical or electronic formats. It may also contain works (e.g., biographies, reviews, criticisms, interpretations, bibliographies, etc.) written by others about the author. This approach is far beyond a simple listing of all items (i.e., different instances of different works by the author) retrieved in response to an author search in existing catalogs. Rather, it would help the user get knowledge of all the works by or about an author arranged in a meaningful order.

In an Author Record, as shown in Figure 1, each work by the author appears once only under its uniform title and the different editions and manifestations of a work are linked (using hypertext linkages, for example) from the uniform title to the Work Record created for collocating them. Pointing to/clicking on each work would bring up editions and manifestations of the desired category or work.

5. The Work Record

There have been some attempts (for example, by Svenonious, 1988; Wiberley, Daugherty, and Danowski, 1995; Fattahi, 1996a, 1996b; Yee and Layne, 1998; and Carlyle, 1999, 2000) to introduce new approaches to reconstruct the bibliographic record for implementing the concept of collocating, integrating and clustering the different items belonging to a bibliographic family. Svenonius (1988) proposes clustering of records representing editions of the same work in displays based on relationships among items. Fattahi (1996a, 1996b) proposes the concept of Super Records for the categorization and integration of documents related to a particular work. This concept has been implemented in the Prototype Catalogue of Super Records (http://wilma.silas.unsw.edu.au/students/rfattahi/super.html). Carlyle (1999: 2) proposes that “one strategy for improving the effectiveness of screen displays in online catalogues is to summarize the contents of sets of retrieved records in one or two screens instead of displaying long lists. This requires the categorization of items related to a particular literary work based on description of the attributes (such as physical format, audience, content description, pictorial elements, usage and language) used for grouping. Yee and Shatford Layne (1998) propose practical ways by which computer software can recognize and organize all of the records that represent one work.

Shakespeare, William, 1564-1616.
Complete works
Selections
Translations
Individual plays (by title):
   All's well that ends well
   As you like it
Hamlet
King Lear
The Winter's Tale

Poems:
Poems in general
Sonnets
Epic poems
Lyric poems

Apocrypha

WORKS ABOUT SHAKESPEARE:
Biographies
Bibliographies
History and criticisms
Concordances
Periodicals (In print and/or Electronic format)

WEBSITES

Figure 1. A sample of Author Record with relevant categorization of works by a classical author

The Work Record would contain the author heading (if applicable) and the uniform title along with a categorization of different editions and manifestations of that work. Each category is linked to the relevant sub-categories linked to records for items available in the collection or on the Internet. It may also display works (e.g., reviews, interpretations, concordances, bibliographies, etc.) about the work being cataloged. Work Records can be created and structured according to the conceptual models developed for entities in the bibliographic family. In this sense, it can be considered a knowledge-based approach to organizing and representing all the instances which relate to a particular work. For example, different editions of Shakespeare's Hamlet which are entered under "Shakespeare, William" (as the main entry heading) as well as different modifications and adaptations based on Hamlet in which "Shakespeare, William" is an added entry can be brought together in a Work Record.

It is possible to retrieve all the editions of a particular work by pointing to/clicking on the relevant category as illustrated below. This type of browsing and searching can be continued to find items and copies belonging to the desired category. Also, moving within different classes of editions and manifestations of the work is very easy with hypertext technique in place.

Shakespeare, William, 1564-1616.

Hamlet
This work includes the following editions/manifestations available in/through this catalogue:
Texts (editions)
Translations
Adaptations or Arrangements (by type of modification)
Changes of Genre (music performances, operas, novelization, etc.)
Versions (by physical form: print, audio-visual, electronic)
Reviews and Criticisms

Figure 2. A sample of Work Record for a classical work
Shakespeare, William, 1564-1616.

Hamlet: Texts (editions): (records 1-10, from 32)

6. The tragic history of Hamlet, prince of Denmark, 1603/ edited by G. B. Harrison, 1924.

Figure 3. A sample of Work Record for editions of a classical work sorted by date

In essence, with the idea of Work Records and Author Records, catalogs resemble bibliographies in terms of their collocating function and the arrangement of entries. What is new here is that with the computer's ability to organize, reorganize and represent entries and to create dynamic links it is possible to construct catalogs demonstrating the relationships between works and their editions and manifestations in a more logical, consistent and knowledge-based way.

6. Implementing the concepts

The concept of Work Record can be implemented through an algorithm that automatically identifies records that refer to the same work and cluster them together. This can be done by pointing to a number of attributes (data elements in bibliographic records) such as uniform title, author, and a term ("Editions", "Adaptations", "Translations", for example) for the category to which a work belongs. Therefore, elements such as uniform heading for authors and uniform title for works are necessary for the clustering of different works by an author and different manifestations of a work respectively. In her research, Carlyle (1999) proposed a practical model for automatic clustering based on types of relationships, including translation, presence of illustrations, etc. The Prototype Catalogue of Super Records constructed by this proved its value in carrying out different functions of the catalog. BOPAC (Bradford OPAC) developed by the Bradford University has implemented a relatively similar concept (http://www.comp.brad.ac.uk/research/databse/bopac2.html).

Research is needed to identify the type of categories within which items related to a work can be grouped together. Also the terms illustrating the types of relationships between and/or among items should be identified, controlled and standardized based on users' preference so that they can be used in all catalogs. Also, how retrievals can be improved using users' feedbacks about the organization and representation of information in IR systems would be another area for research.

All such new approaches need librarians' knowledge. Librarians select, analyze, evaluate, and organize relevant information and thus create new knowledge. In this respect and based on their wealth of experiences developed over the history of knowledge organization and representation, catalogers and indexers have a lot to offer the new environment. Librarians are knowledge workers making
the transfer of information into knowledge easier and practical.

References


Devika P. Madalli and A.R.D. Prasad
Documentation Research and Training Centre
Indian Statistical Institute, Bangalore, India

VYASA: A Knowledge Representation System for Automatic Maintenance of Analytico-Synthetic Scheme

Abstract: The present work deals with building a computational model for maintenance of a classification schedule. It has been hypothesized that the analytico-synthetic approach is quite amenable to the tools and techniques of Artificial Intelligence. To achieve the objective of building an automatic system for management of classification schemes, 'Vyasa', has been developed using AI tools and techniques.

1. Introduction
The present work is a sequel to the earlier work, where the application of Artificial Intelligence tools and techniques were used in the verbal plane i.e. subject indexing (Aptagiri, et al., 1995; Prasad, 1993) and later extended to notational plane where the system could assign automatically the class numbers to given expressive titles (Panigrahi, 1996). The present paper is based on the work carried out to update the 7th edition of Colon Classification (Madalli, 2000).

2. Need for the Study
A scheme for classification cannot be exhaustive at any point in time, because knowledge is dynamic and growing and so also are the different subject ideas and concepts. All these need to be represented and accommodated in a scheme for classification. This warrants regular updating of schedules to keep up with the changes in the universe of subjects. The various reasons, which necessitate such updates, may be:
• To accommodate new concepts/subjects
• To delete old/obsolete concepts/subjects
• To accommodate the change in denotation of existing concepts/subjects
• Structural changes in subjects

The objective of the present work is to design a computational model to automate the updation of analytico-synthetic schemes for classification, in particular Colon Classification. The main objective or goal of the present work can be further divided into the following sub-goals or secondary objectives:
1. The system should be capable of processing expressive titles of documents.
2. The system should be capable of representing the classification schedules intuitively.
3. The system should be capable of incorporating the rules of classification for effective inference mechanism and construction of class numbers.
4. The rule based expert system component should be able to assist in accommodation of new terminology to deal with the work at a practical level.

3. Methodology

Frame based knowledge representation allows the representation of schedules in a straightforward way. It is to be noted that frame based knowledge representation enables building hierarchical structure of facts, where the relations of the concepts to other concepts or isolate ideas can be inherited by facts at a lower level from the higher levels. This is exactly in accordance with the purpose and structure of analytico-synthetic classification schemes.

Rule Based expert system facilitates the management of schedules for functions like its updation. The principles and postulates stipulated for Colon Classification are explicit and make it amenable to be expressed as rules for the inference engine. Updation of schedules in a scheme for classification involves the analysis of the expressive titles into different facets. Work in computational linguistics focuses on analysis of natural language sentences. Natural Language Processing allows the design of parsers and recognizers that scan a given expressive title and cull out the significant terms and other clues like parts of speech of the terms.

The present work concentrates on the areas highlighted in the preceding discussion. It tackles the problems related to the updation of an analytico-synthetic classification scheme. This is achieved through building a knowledge base through frame-based representation of the schedules. Rules for management of the Colon Classification (CC) schedules are incorporated in the rule-based module. A natural language parser is used for analyzing the expressive titles into noun phrases. PROLOG is used for writing the source code.

4. The Updation Paradigm

The steps in the updation of the schedules are incorporated into an inferencing mechanism. Firstly, the system checks for the existence of the terms, which are output noun phrases of the NLP component (which uses a DCG parser) in the system knowledge base. For the terms which are represented in the system knowledge base, the system assigns the CC notation straight away. However, if a term is not found in the knowledge base, the system prompts them as ‘unknown’ terms.

4.1. Puff Words

The system first attempts to check whether the unknown term is to be treated as a puff term. However, a term may be an insignificant term or puff word in one instance, but may be a significant term in some other instance. To ascertain whether a term is puff word or not the system asks the question: *Is the unknown term a puff word or significant word in the given expressive title?*

Illustration:

1. Title: [an, introduction, to, human, digestive, system]
   
   Noun Phrases: introduction
   
   human ^ digestive ^ system

   Unknown Term: human
Do you consider human a puff word in the present Context? (y/n):

Figure 1: Work flow in Vyasa

4.2. Synonymous Terms

If the term is not a puff word, the system proceeds to ascertain whether the new term has a synonymous term by asking: 'Do you think the word/phrase is a synonymous term to another term?' If the answer is 'yes', the system will ask the present phrase to be entered and the standard term. Once the standard term is given it updates the knowledge base by adding a fact similar in syntax to the following:

\[ \text{value(present\_phrase, use, standard\_term).} \]

4.3. Totally New Terms

In the instance that the unknown term recognized by the system is neither a puff word nor a term synonymous to an existing one, the system treats it as a new term and a claimant for a position in the schedules and the existing schedules have to be updated. With regard to updating the schedules, the following three possibilities should be taken into consideration:

a. First study whether the schedule is dynamic or static. For example, the Time schedule is the most static and exhaustive. This seldom warrants any kind of updation.

b. The other possibility is that the concept is same and the term for it also exists but there may be a structural change. This happens in schedules, which are
dynamic in nature. For instance, the Space schedule is relatively dynamic. For example, for Ukraine the BT was Russia; Now, the BT is Europe. In such cases, the system knowledge base has to be updated manually to accommodate the changes.

c. Totally new terminology. In this case, the system performs Inferencing to determine the position of the new term in the schedules. The system tackles the issue at the following levels:

- Context based inferencing
- Linguistic clues
- Interactive inferencing
- Notation assignment

4.3.1. Context Based Inferencing

The system attempts to take clues based on the context in which the unknown terms occur. The process of context based inferencing is carried out in the following two ways:

1. When the system comes across an unknown term, it presents information about the known terms of the input, that is, the expressive title. It displays the basic subject, category, isolate number and symbol for the known terms in the input. Refer illustration in section 4.3.4. The output gives an indication of the basic subject of the unknown words. In a way the system is suggesting the basic subject of the unknown term under the presumption that if other terms belong to one basic subject there is a likelihood that the unknown term also may belong to the same basic subject. Based on the context the system tries to guess the basic subject. Since the other terms in the title belong to the basic subject, ‘L’ — ‘Medicine’, the present unknown term may also belong to the same. However, this may not be the case if the expressive title contains phase relations or warrants the use of subject device.

2. In addition, the system tries to find out if the unknown term occurs any where else in the knowledge base. The term may exist as part of another in another context. The system after recognizing an unknown term, presents all the phrases where the term appeared. At this stage it prompts ‘The unknown word occurs in the following compound words,’ and presents all the compound words in the above mentioned illustration.

*The Unknown Word human appears in the following compound words: human ^ embryology*

In the above example only one compound term is found. In some cases more than one term may be found. This approach is used to give a clue so that the user asserts a relation between unknown term and one or none of the compound terms displayed. The system displays the compound terms, where the present unknown word is a component by using the following linguistic clues.

4.3.2. Linguistic Clues

The system tries to find out whether the term is a noun or an adjective from the lexical entries. It tries to guess the noun form or the adjectival form of a given term and looks in the knowledge base if the term exists. If it fails, the system attempts to fabricate terms having a maximum of four words where the unknown word plays the role of an adjective and then as a noun. The same can be implemented for a term of any number of words as a recursive procedure is used. In a four-worded term the noun may be the fourth term, or the third word in the three-worded term or the
second in a two-worded term.

Two unknown predicates are used, the first one attempts to look in the lexicon, whether the unknown term is a noun. If it is not found, the system looks whether any variant (spelling or plural) of the unknown term exists in the lexicon. If the term is found to be a noun, the system looks for four-, three-, two-worded compound terms with the present term as the noun.

For the following system generated terms (where T1, T2, T3 are adjectives and Variables), the system will search the knowledge base. If a match is found, it will be displayed in the output, so that user can take a decision whether one of the compound terms found is related to the present context.

\[ T_1^A T_2^A T_3^A \text{noun} \]
\[ T_1^A T_2^A \text{noun} \]
\[ T_1^A \text{noun} \]

Similar approach will be followed to generate compound terms where the presently unknown word could be an adjective.

4.3.3. Interactive Inferencing

If all the above attempts fail, then the system asks for the broader term to the unknown term. The system then proceeds to write the relevant facts for the unknown term into the system knowledge base. The system then presents the schedule where the BT occurs, along with all the NTs under it and prompts, 'The following are other Nts under the Bt: ' .....' the nts predicate finds the narrower terms for a given BT.

4.3.4. Notation Assignment

Notational Plane of Colon Classification: The number of ideas or isolate ideas, are numerous in the idea plane. The verbal plane, be it of any language falls short in coining terms to express all the ideas exhaustively. Even though only technical terminology is used in classification schedules, still there is a great deal of expressiveness through words. But the notational plane falls very short and faces great challenges in trying to express the concepts in the verbal plane, as any notation is an artificial language. Ranganathan proposes the concept of co-extensiveness in notational plane where co-extensiveness is defined as 'The representation in a class number of the measure of incidence of each of the relevant characteristics, of the subject embodied in the document classified' (Ranganathan, 1967).

Generation of sector notation in Vyasa. Vyasa has incorporated a module to generate the sector notation as demonstrated by Ranganathan. The total number of co-ordinate classes including the Packet Notation is 1,166. Once all the numbers are generated, they are stored in a b-tree in sorted order.

Notation Assignment: interpolation and extrapolation In continuation of the flow of work of Vyasa in the automatic updation of schedules, the system asks the user to supply the broader term of the unknown term. Once the broader term is supplied the system displays the array/s under it. The user can then choose the position of the unknown term in that array. Here the Principles of Helpful sequence act as guides to the process of decision making.

A find nts predicate is used to suggest/assign the appropriate isolate number. Once the system displays the array of the Nts beneath a given Bt, then user selects the position of the unknown term by the isolate number of one of the displayed isolates. The system then asks 'Do you want a preceding number?' for which if the assertion is 'y' -'yes' then it says 'Shall I assign the number -- for the term ----' for
which again if the assertion is 'y'- 'yes' then the system assigns that number to it.

The preceding number for a given number is obtained from the tree generated by the zones predicates. To obtain the preceding number, the system uses the following predicate. Similarly, the system asks whether a number succeeding the one chosen should be assigned and proceeds in a similar manner to the above to assign the number.

Intuitive learning is an important feature of expert system. Vyasa incorporates the new terminology into the system knowledge base, once it gathers the relevant information regarding the new/unknown terms. Thus it updates the system knowledge base so that those terms are automatically assigned the number next they appear in the input titles.

Illustration
Title: [diagnosis, of, aids, disease]
Noun Phrases: diagnosis aids^ disease aids
unknown term aids

**** INFORMATION ABOUT THE TERMS FOUND ****
Term: disease Term : diagnosis
Basic: L Basic: L
Cat: mp Cat: r_energy
No: 4 No : 3

Do you consider aids a puff word in the present Context? (y/n): n
Do you think the word/phrase aids is a synonymous term to another term? (y/n): n

Provide the BT of the term - veneral^disease
The following Fact is added to the knowledge Base
value (aids, part_of, veneral^disease).
The following are Other Nts under the Bt: veneral^disease

4221 Syphilis
4223 conhomless
4224 sofisorie

Do you want a preceding number: n Do you want a succeeding number: y
Enter the number: 4224 Succeeding No: 4225
Shall I assign the No: 4225 the number to the term: aids (y/n): y
The following Fact is added to the knowledge Base
value (aids, number, 4225).

L;1;4225:3

5. Conclusion
Vyasa has successfully demonstrated the integrated use of frame and rule based knowledge based techniques in maintenance of analytico-synthetic classification schemes. In the process it brings out the features that make the adaptation of AI techniques feasible. The schedules of the various subjects can be intuitively represented in frames and easily manipulated by using the rule-based module. Here Dr. Ranganathan’s meticulous rules helped tremendously as they could be interpreted and implemented without any ambiguity.

References
Images and Words

Abstract: The image as a new indexing resource to be integrated to the information systems, must be studied with the final purpose of representing its contents according to the special characteristics it exhibits. Differences between the traditional documents and the singularities of digital audiovisual information that is substituting analogical information, are being researched. Another aspect being examined is the re-evaluation of the thesaurus and the convenience of its use in the new informative environment. It has been proposed an image analysis extrapolated from the Panofsky’s proposal to analyse art works and, therefore, foresee the representation of the different elements necessary for the audiovisual document analysis in the thesaurus to be built.

1. Introduction

The human being is exposed to a constant exercise with unreal images, taking into account television, computers, publicity at the streets and in the movies. Sixteen out of the twenty four hours of the day, he’s submitted to images that try to reproduce the real world. When sleeping there’s also image activity; dreams occur with images and imagination is accompanied by a representation with these ones: we imagine situations, solutions and actions. There’s even the proposal that man’s memory is stored in our brain in the form of images. The evolution of the intellect, and together with it, that of the visual-perceptive process, leads the man to develop an exceptional skill: that one of generating representations that “imitate” objects and scenes of the environment. (Yankelevich, 1993, 15-31)

The traditional photograph-illustrated document is being substituted in part by the audiovisual one, that nowadays has been transformed from analogical to digital systems, converting images, sound, music and voice in manipulable and feasible-to-transform data and intercalating them to generate another type of product, different from the original. Videos, movies, sound, texts and graphics are digital substances expressed in data. This characteristic of the digital information takes us back to many menhires and dolmens from the Stone Age, used by the Celtic druids to register their formulas and enchantings (it is said that the stones they used came from the Stone Age but the most accepted tendency affirms that these inscriptions dated from the Christian Era); also to take notes employing the alphabet known as Ogham, that originated from Ogma, the Celtic God of Writing. (Rosaspini, 1998, p. 180)

Davis confronts digital information with the Ogham stone because the traces of this writing are like the digital coding of the ancient Irish language, due to the fact that this alphabet uses a binary code, being expressed the letters and syllables by grouping vertical bars. However, the difference is that digital information is coded with zeros and ones, we cannot read directly from them as from the Ogham
code written in the stone that, actually, cannot be read either, unless we know the meanings of the glyphs. (Davis, 1997, p. 292)

"Digital technologies change the creation and information movement into a unique substance that is infinitely transformable. Film, vinyl, magnetic tape, paper, photosensitive paper, ink, graphite, paintings, all of them transformed into digital domain create a new medium different from any other thing previously invented. The digital substance can be transformed into any other thing by means of the computer. The audio data can be transformed into representative images. All the media are transformed into data types. We return to the Ogham stone, that is, coding all the ideas in a binary inscription". (Davis, 1997, p. 294) (translated from Spanish)

Few years have passed after the merging of the audiovisual and digital information; the literature generated about its organization in the librarianship field is not abundant, according to the originated expectation concerning a technology that appears to be very simple but that requires a greater study, not only about the way to index audiovisual documents but also about how to establish associations between the different contents in order to represent it and even about the way to design and create the digital document icons, issues that will not be treated in this paper.

2. Images Treatment in the Documentation

Images treatment in the information transmission processes has provoked a new paradigm in documentation, that leads towards the solution of complex methods and less frequent use of the natural language, for content analysis and image representation as object of information.

Documentation has followed with great interest not only the evolution of information processing, traditionally destined to the service of the information units, but it has also evolved together with the merging of the new sources and supports for the information transfer such as the image and the diverse media that carry it. This transformation has been triggered by the treatment of images within the calculation systems and services that has allowed the establishment of the image over the word in audiovisual documents for teaching, commerce and research.

The evolution that has affected teaching and knowledge transmission in recent times, leads towards distance methods, a field in which image has been reassessed as a resource with a high level of information density, and that requires with increasing urgency the creation of codes to represent it and help to decode its content associating it to the new information needs of the users, teachers or intellectual workers.

The librarianship's role in the synthesis and analysis of information has been more appreciated due to the fact that it has not been limited to the description of form and content of traditional documents; on the other hand, according to this new need of interpreting contents in a world saturated with images, the librarianship has had to lean on thesaurus techniques that help him normalize to a certain extent images and contents associated in the web, so that the librarianship has become a sort of cartographer to the cyberspace navigators.

The abundance of images and the necessity to classify and organize them, makes necessary the creation of controlled languages with a hierarchical structure to represent from the most general to the most specific topics, relating them among
themselves and with other similar issues. This task has become a necessary work for information analysts, associated to the new knowledge production trends. Linguistic mediation through a documentary thesaurus with representations, including the different aspects of images’ content allows the indexing for further recovery of the digital image and audiovisual documents in general.

This paper originates from the premise of the thesaurus as a system that is much closer interrelated with the digital document as another system in itself, than natural language (that is without underestimating this late one as a searching option). The thesaurus and the digital document are interrelated according to the recovery of subject matters within an information system to facilitate the access to and the use of the digital document as such.

Information systems are part of the productive cycle of knowledge, the task of which is to transfer these information systems for social use. Registers of informative documents conform the information system and are structured in a way that they not only can be consulted one by one but recovered through access points strategically selected; they can also group common elements to several registers through the activation of document categories containing the same concepts. According to this, it can be introduced Lafuente’s definition about the operations of information categorization concerning documents, which presents as the generalization of a concept abstraction, related to all the documents containing it, “...the results of the analysis of document groups that are expressed in the form of document categories, that is to say, in abstract concepts defining their common properties and their most general relationships. These categories are the result of an abstraction that generalizes the particular or singular aspects of the documents produced and used by a community.” (Lafuente, 2001, p.163) (translated from Spanish)

The categorization and grouping properties of the documents analysed in the systems help the user solve specific problems of information. Common issues that can be found within a system of information are organized by a thematic classification, through a vocabulary control that helps the analyst as well as the user to conceptualise the issues contained.

3. Indexing of Images

Santos Martínez proposes a model of iconological analysis for the cinematographic short story that seems to be also adequate to be extrapolated to the audiovisual information. (Santos Martínez, 2001, p. 2) The principles it uses come from iconology, which is the branch of history that deals with the theme or meaning of the works of arts, disregarding the form. They also come from iconography that studies the form. (Panofsky, 1970, p. 41)

Panofsky’s contribution consists in the classification of the different meanings present in a work of art; in order to do this he relates us common episodes of man’s life and classifies them with the same criteria that he uses to classify the work of art. According to this, it can be explained the extrapolation towards the audiovisual information that we need to do “...by transferring the results of this daily life analysis to any work of art, we can distinguish the three same strata in its theme or meaning.” (Panofsky, 1970, p. 39)

The primary or natural meaning, subdivided in factual and expressive is the one apprehended by identifying the pure forms, that is, specific line and color
configurations or specific masses of wood or stone with peculiar shapes, as representation of natural objects such as human beings, animals, plants, houses, tools, etc. The transmission of meanings by means of the form constitutes a pre-iconographic description of the work of art. Let's set an example, in the film about the sinking of the Titanic, the primary meaning of the scenes picturing the tragedy would have to be related to a sinking ship, the sea at night and human beings drowning.

The secondary or conventional matter is connected with themes or concepts. The motives acknowledged as transmitters of a conventional meaning can receive the name of narrations or allegories. The identification of such images, narrations and allegories constitutes the domain of what is commonly designated as "iconography". When we talk about "matter as opposed to form", the reference is to the field of the secondary or conventional matter, in other words, to the world of specific themes or concepts manifested in images, narrations and allegories, in opposition to the field of the primary or natural matter manifested in objects. Returning to the former example, the secondary or conventional matter is the sinking of a huge ship occurring during the night, that didn't have enough lifeboats or lifejackets for all the passengers.

The content or intrinsic meaning is recognized above the underlying principles that are manifested through an attitude or lifestyle, condensed in a play. These principles are revealed through the composition methods as well as the iconographic meaning. The pure forms, motives, images, narrations, allegories are recognized as manifestations of underlying principles, interpreting all these elements as symbolic. The discovery and interpretation of these symbolic values (that are many times unknown by the creator of the work and that may even differ from what it was intended to express) is the object of what Panofsky designates as iconology. In our example about the Titanic, the content or intrinsic meaning referring to the sinking scene prompted by the intensity of the music, the sound, the actors' physical appearance, their anguished faces, represents the impotence in front of nature, the immensity of the ocean, the sense of prioritise profits over safeness, etc.

Therefore, we can find three distinguished levels in an audiovisual: the pre-iconographic description, the iconographic analysis and the iconological interpretation.

In order to organize a digital information system, the audiovisual documents must be represented from the viewpoint of the theme or event dealt with alongside the document, the same as in the traditionally-printed information systems. But besides all that, they must represent very punctual objects, although their appearing is very ephemeral (as in the case of characters, institutions or very specific information), and whether this is related or not to the main theme of the original or main digital document. In a digital information system, the presence of the descriptors that represent it is, undoubtedly, more abundant according to quantity than in the traditionally printed systems. (Burke, 1999, 96)

The transmitted knowledge by means of the audiovisual document is a human cultural product and the word is used in order to transmit it, due to the fact that the word gives it the meaning, the precision and the clarity even to silent images, such as photographies. In a documentary system, the words that represent the meanings of the included documents are represented in the descriptors selected to index the document and extracted from a thesaurus. In general terms, the
objective of the content description is to determine the field of knowledge, classify
the different meanings that conform it, express them in concrete terms and without
ambiguities, translate such terms to the most appropriate language expressions, all
of this with the final purpose of:

- Acknowledge the document object of analysis in a whole mass of
documents.
- Represent the knowledge and learnings implicit in the document through a
classification.
- Substitute the document.

Until recent years, the predominant paradigm was the linguistic
communication for information, teaching and even for entertainment (Enser, 1995,
127), but in the multimedia digital documents the image constitutes the speech
central element. It occurs the same, for example, as with a surgery procedure
recorded for medicine students or the internal exploration of the human body.
Images constitute a language analogous to any other the human being deals with,
such as the verbal or the musical one. Despite all this, the image must be
accompanied of an explanation of what is being seen, in order to guide the brain
and the sight. “...and, however, the image is not the spoken tongue of our children
because it lacks syntax and grammar. An image is neither true nor false, neither
contradictory nor impossible. As long as it is not an argumentation, it is not
refutable. The codes that it can or cannot mobilize are just readings and
interpretations.” (Debray, 1994, 53)

The image must be represented in its contents in order to locate it among a
series of documents with images and can only be represented by indexing terms
belonging to the written or spoken document, from where key words for indexing
can be extracted. If a written document deals, for instance, with Greek columns,
only one term would probably be enough to represent the theme. In an information
system containing the image of a Corinthian column, although the spoken text
accompanying it is general, if the system policies constitute the precision of artistic
elements, the column of the image should be described; moreover, it’s even
probably required to add the type of acanthus- shaped leaf decorating its capital.
Another variable would be added to this image if, for instance, next to this Greek
column, the singer Maria Callas would appear. According to the content description
she should be represented by an identifier, in this case, a character one.

Abstract or concrete ideas concerning any topic can be expressed in a
document, even when it refers to more concrete situations represented in an image.
Besides the theme, in the image it is important the shape, the color, the epoch and
the details (Chen y Rasmussen, 1999, p. 295). To exemplify this, we can mention a
written document about the landscape of a geographic area and an image
representing that landscape. In an information system, the written document will be
well represented by a descriptor that mentions the area, the type of landscape, the
season of the year, the time of the day (it can be sunrise or sunset), the most
notorious species of trees, etc.

Digital information includes image constructions that must be re-used many
times due to its high obtainment cost in order to record them; sometimes because
they constitute events or characters that could not be recorded again; or due to their
distance in reference to the physical or temporary environment, or the immediate
need. Some other times it is due to the difficulty to record them given the fact that, occasionally, even images created as digital objects consume edition work hours.

It is important to review the forms, grammatical categories, grammatical accidents and descriptors' thesaurus relationships to be able to understand not only their adequacy to the structure within a digital environment, but also the changes introducing those elements in a system with the capacity of integrating images and sound. Changes may not be substantial, nevertheless their presence is noticed when beginning the debate about each of their aspects separately.

The indexer abstracts the contents to be represented through a word synthesis, by means of a visual-perspective and analytical process of the document images and their explanation by means of the sound. This interlaced parallel among the natural environment perception, its replica through images and its representation by words, is carried out inside a dynamic information exchange in the community (Yankelevich, 1993, p. 18). The set of participant individuals in this relationship learns and shares the new forms of observing reality and the capacity of expression through a thesaurus that includes the adequate denominations for objects, contents and meanings of the audiovisual content.

4. Final Considerations

Description as well as representation of contents in the digitalized audiovisual information constitute basic elements for its broadcasting and recovery. It's necessary to study the analysis and synthesis process to be able to translate the knowledge they contain and incorporate it to the information systems. The Panofky's method to examine the audiovisual information constitutes a proposal to be taken into account.

Thesaurus constitute a linguistic tool structured in descriptors' categories and are adequate to represent the audiovisual digital information, if the different aspects of their content are incorporated to them. The representation of images demands the use of a greater amount of indexing expressions due to the fact that forms, conventions, contents, characters and institutions must be interpreted.

Words are valid to represent images and this is the adequate way to do it because, through language, the intellect abstracts what is observed in three dimensions and represents it.

References


Abstract: This paper focuses on the need for knowledge organization (KO) tools, such as library classifications, thesauri and subject heading systems, to be fully disclosed and available in the open network environment. The authors look at the place and value of traditional library knowledge organization tools in relation to the technical environment and expectations of the Semantic Web. Future requirements in this context are explored, stressing the need for KO systems to support semantic interoperability. In order to be fully shareable KO tools need to be reframed and reshaped in terms of conceptual and data models. The authors suggest that some useful approaches to this already exist in methodological and technical developments within the fields of ontology modelling and lexicographic and terminological data interchange.

1. Semantic interoperability, the WWW and knowledge representation (KR)

The evolution of computer networks has stressed the need for data and information interoperability, i.e., for having data and information assets reusable across distributed and heterogeneous systems. According to Amit Sheth (Sheth, 1999), we are now in the third generation of interoperable systems where the concerns are mostly focused on information and knowledge, emphasizing semantic interoperability at a level higher than that of previous developments. Before the expansion of the Internet, interoperability was concerned mainly with intersystems communication and agreements on data syntax and structure for communities of systems (multidatabases, federated databases or federated systems). The WWW brought a new dimension to fundamental concepts such as distribution – from the enterprise-wide space to the global space – and heterogeneity, implying changes in systems paradigms that are far more complex than just a matter of scale. Another fundamental aspect that also became more complex to deal with is autonomy – a systems requirement that has to be balanced with the increased demands in network interoperability. Such demands have influenced turning points in systems’ architecture and design, highlighting the trend for ‘composability’ of solutions, in which components tend to be system independent, adaptable, extendable and reusable. This is true both for software engineering and information design and data modelling.

The trend described by Sheth – from system, syntax, structure to semantics – is well illustrated by the semantic Web movement, and all developments around it. This is especially true of XML, as an independent language to structure resources, and of RDF, as an XML specification to convey machine-understandable
representations of resource descriptions, including content description, and metadata modelling, to build such representations. On top of that, the use of common or shared formal languages (i.e. controlled vocabularies) to convey explicit and shareable representations, as well as of ontologies to support them, became part of what is now understood as the architecture for the future Web (Berners-Lee, Hendler and Lassila, 2001).

Developments around formal languages, ontologies and vocabularies touch three major fields: the field of computing and IT, mostly devoted to raising methods and methodological tools for building controlled knowledge representations (KR), the field of terminology and lexicography, and the field of knowledge organization, where classification systems, thesauri and other controlled vocabularies for information retrieval are largely produced. Most of the theoretical and methodological research around ontologies has flourished and intensified in the last decade to semantically support expert systems (Sheth, 1999; Kashyap and Sheth, 2000). It encompasses diversified levels of approach, from the theory of knowledge representation (Sowa, 2000), to applied fields such as knowledge bases, new methods of software engineering (Gruber, 1993; Guarino, 1998, Guarino and Welty, 2000), or information brokering based on metadata for knowledge domains (Kashyap and Sheth, 2000). Because the use of ontologies in these approaches always imply some form of formalized logic, they developed mostly from the theoretical background of artificial intelligence (AI). Nevertheless, the methods and tools to formalize and structure such ontologies have evolved in ways that are becoming more amenable to use by non AI experts. Therefore, they provide opportunity for exploring new principles and solutions that can benefit renewal in related areas such as knowledge organization for information retrieval.

2. Disclosing knowledge organization (KO) tools in the network

Not only ontologies have emerged “from academic obscurity into mainstream business and practice on the Web”, as noted by McGuinness (McGuinness, 2001), but also the Web environment has raised the importance and value of existing KO tools, such as classifications, thesauri, taxonomies, subject-heading systems etc. These are now claimed to have an enormous potential, not only in being applied to the self description of individual Web documents but also in supporting search and retrieval services provided by agencies other than libraries. This is evident from both the current literature on the matter and from the number of projects, agencies and metadata schemes, etc. that recommend or refer to the use of traditional library KO tools. Apparently, these are ready to use and well credited because they are professionally prepared resources and because they reflect literary warrant. Nevertheless, they show actual constraints that are not minor drawbacks to the objectives of KO tools being widely shared in the network. Such constraints are analysed from several major perspectives in the sections following.

2.1 Explicit ontological frameworks

Although KO tools represent intellectual constructions at its best, backed by conceptual principles, these are often not explicitly conveyed, i.e., in forms that could speak for the system as a whole. This addresses information to clearly identify a system’s domain, its boundaries, structure and major changes during the course of its evolution, categories of concepts and principles governing their relationships, policies regarding relationships to other KO systems, etc. In fact,
what is usually available in the first plan, even for the professional user, is the resulting product - the structured vocabulary of the thesaurus, the subject heading list, the classification schedule - not the underlying philosophy, principles and policies, which in many cases happen to be documented long after the corpus exists. In the larger context of standards, the few international guidelines available in the field (such as ISO 2788 and 5964, for the construction of monolingual and multilingual thesauri, respectively) have long been recognized as insufficient. They are very basic, about three decades old, do not provide for every kind of KO system and do not have the strength to create an agreed ontology for the universe of discourse in the KO field. It is even a paradox that the field of knowledge organization still lacks a sound basis for conceptual and terminological consensus, as it is true that actual tools are still biased by their historical and local technical traditions.

2.2. Open network availability

As resources in their own right, i.e. apart from being provided as discrete elements in bibliographic systems, KO tools only recently became available on the network, although since the nineties some of them have been published in electronic form, e.g., in CDROM as an end-product or just as structured electronic files to be handled by a database system. Besides being available on the network, KO tools need to be designed for different users and usages, including the possibility of being accessed and used by automated agents. For instance, it is important to consider KO tools as identified and maintained persistent namespaces, and to include a structure to allow external entities to link and refer to any of their constituent elements. This is a requisite that fits into the framework of the Web semantic architecture and that may in turn result in modelling requirements at the level of data model or systems architecture.

2.3. Data sources shareability

Data sources shareability refers to the level of portability of data content among different systems, for the purpose of reuse. Several levels of pre-requisites are needed for this, which will be presented starting at the lowest level. Data representation language is the basic level and should not be particular of a given system, group of systems, or platform. This naturally points to XML, if not as the language for holding data in a system, then at least as the possibility of transforming data into XML at the export level. The next level is data structure - i.e., the data representation aspects that determine levels of syntactical compatibility (machine-readable aspects of data, such as data element definition, also referred to as formatting) and logical compatibility, referring to the main components of the underlying model (type and definition of entities, attributes, relations). Data structure should provide for data interchangeability among systems deemed relevant to each other, at least at a minimum level. In practical terms, this would make data transportable without loss of significant content and features, and ensure that it is reusable without major conceptual conflict. These aspects require some level of conformity to common representation standards and declarative tools about data components, features and options particular to each system, at a meta-meta level. With regard to data sources shareability, KO tools used and/or maintained by libraries have improved little by the efforts of library automation. So far, the automation of KO tools assumes one of two forms: either an authority MARC file, designed mainly as a subsidiary management resource internal to a given bibliographic system, or an independent database, primarily oriented for editorial
and publishing (until recently, mostly printed publishing) objectives.

In the first case – MARC authority files, either for subject alphabetical or classification systems - the underlying functional and data models derived mostly from bibliographic management requirements and did not evolve according to other purposes that could be additionally envisaged for the KO tool as an independent resource. For example, MARC authority files have not been used to display the KO tool itself independently of bibliographic data. While in this case the solutions provide for data shareability within the community of MARC systems, such data structure is difficult to share with non-MARC systems, as it is. Simply raising a non-MARC format from scratch would not solve the problem, because the primary community of KO tools are libraries and they all talk MARC, so MARC formats will continue to evolve. One example is the **UNIMARC for Classification Format**, recently issued by IFLA, whose objectives are twofold: to support authority control functions in bibliographic systems and to serve as a common format used by publishers to deliver classification data. Additionally, it can also serve as the logical format used in classification management systems. Nevertheless, one has to recognize that the model underlying MARC for classification did not explore what were the additional requirements for that purpose.

In the second case – KO systems managed independently of library systems – there are no standards or common data models whatsoever. This is the case with most solutions used to manage thesauri and classification schemes, meaning two things: first, that the shareability of such data in a machine-readable form is not even part of the requirements of such systems; and second, that efforts in the modelling of such KO tools and supporting systems remain isolated and idiosyncratic. They do not aggregate the modelling needs of a community and their data structures remain mostly invisible. Yet, in cases where electronic files are provided, they usually imply a replication of the solution used by the provider, and the knowledge particular to its format.

### 2.4. Data content usability

There are two different aspects regarding data content usability. First, the clarity of intended meaning of data element instances. This refers to the level of semantic expressiveness provided, by means of definitions, attributes and relations included in the actual content of KO tools, and also to the way they are conveyed to the user. Second, the aspect of alternative usages, i.e., different views of the same data content for different kinds of users and in contexts other than those where the tool originated or to which it was primarily designed. In both respects more could be achieved in terms of usability, if the functional requirements were revised on the assumption that the tool is to serve a wider community of users who are not only end-users – in the sense of bibliographic searchers, or information seekers in general - but also metadata information providers. On one hand, KO tools could consider expanding content by interlinking with other network resources, e.g. referring to other KO tools, or to major resources about a given concept, or instances of a class. On the other hand, data presentation models could be re-analysed in order to improve methods of content and context display, for example by overcoming the traditional use of “condensed” conventions (like symbols and abbreviations) for which the major justification in the past was the “economy of space” imposed by manual products. Today’s systems have no reason to live by old assumptions and limitations, if the target usages are changing and if the technical means available make changes affordable.
2.5. The centrality of data modelling

From all the above perspectives one can easily derive many aspects that would not only modify the model of the traditional KO tools, but could also lead to an assessment of data models. Data models, and standards that support them, are critical for shareability as they address the real technical and semantic conditions without which network availability and content usability are of limited use. The centrality of data modelling comes from its nuclear position in the design chain: it conveys major elements of the conceptual model (elements upon which information is designed), it interprets requirements and it provides support for functionalities. The definition of data standards is the most visible result of data modelling, especially in the case of a new standard. But the overall implied activity goes beyond that, it is not merely a stage in development, it is a level of management that should be continuously exercised, because systems evolve with changes in information. The emphasis on data models is therefore justified, as a critical piece of the information architecture either in individual systems, multi-systems or open information space design. KO managers should begin to think about modelling at large, i.e. expanding the data model beyond the local, individual or specific community’s system requirements, either in informational or functional terms, in order to enhance their network value.

3. KO modelling for the future

The conceptual framework for KO tools has much in common with the conceptual framework of an ontology developed for knowledge-based systems. Languages for modelling and representing ontologies have been developed for more than two decades, but not until the advent of the Web was their full potential in information exchange and communication revealed. Also, Web technology itself has helped to advance ontology-modelling languages, through open standards such as XML and RDF.

Before the web, ontology modelling languages were normally characterised by high computational complexity and logic-based formats with specific syntax and a function of mapping ontology from and to computer languages. A typical example is KIF - Knowledge Interchange Language. The new generation of Web-oriented ontology languages are now being developed. DAML - The DARPA Agent Markup Languages, OIL - Ontology Inference Layer and SHOE - Simple HTML Ontology Extension are examples of such languages aiming to achieve the status of standards. They maintain the expressive power of a logic-based language while also being system independent and Web compatible through their adherence to XML and RDF. Even languages like UML (Unified Modelling Language), which is primarily a general modelling language for object-oriented analysis and design, is successfully applied in ontology modelling when used with RDF syntax (Cranfield, 2001). The fact that tools for ontology modelling are now made available and understandable even for those without programming and AI expertise opens up new opportunities for the application of ontologies (Ying, 2000). The framework for data analysis and conceptual modelling that is offered by these languages is an important reference point for everyone involved in KO data modelling, as they offer methods for strict analysis of data content, data logic, class definition, property inheritance, etc.

Besides ontological modelling languages, another area of web-based
standards that is of particular interest to the KO field is that of developments supporting terminological exchange. These can also become a source of useful approaches in modelling-for-exchange. Internet multilingual demands have encouraged numerous developments in automatic translation and stimulated research in the field of lexicography and multilingual terminological exchange. Among the first standards for terminology exchange was MARTIF - *Machine Readable Terminology Interchange Format* (ISO 12200), designed as an SGML based format for human-oriented terminological and lexical databases. More recent is OLIF - *Open Lexicon Interchange Format*, an XML-based standard that builds on MARTIF, especially improving the aspect of machine readability of data for machine translation. Building on these two standards, which are primarily focused on lexical data, a new standard is being developed: TMF - *Terminological Markup Framework* (ISO 16642). This one is more networks oriented and includes features for conceptual and ontological aspects of terminological data (Romary, 2001).

These developments from the areas of AI and lexicography could offer a useful background for the reshaping of knowledge organization tools in terms of modelling activities. The aforementioned standards and terminological formatting languages could offer more logical and structured methods for representing linguistic aspects of KO tools in order to enhance them. An initiative that appears to be in line with this trend is Voc-ML - *Vocabulary Markup Language*, raised in the context of NKOS - *Networked Knowledge Organization Systems* promising tsupport to KO tool exchange. However, it is evident that more joint effort from the KO community is required to make advances toward more interoperable and shareable KO models.

5. Conclusions

This paper analyses the current state of knowledge organization (KO) tools such as classification schemes, thesauri and authority files, focusing on the strengths and weaknesses of their electronic handling and exploitation in the Web environment. The importance of semantic interoperability stresses the potential of KO information assets in the framework of the Semantic Web, calling for improvement and change in the network shareability of KO tools. This highlights the lack of common methods and standards to support the current needs of cross-domain and multi-system usability of KO data. Existing constraints in the way KO systems are currently managed and made available indicate the need for their significant improvement at the level of modelling, especially in what concerns the modelling of data structures, as a fundamental condition to disclose KO resources over the network. In this respect, an interdisciplinary approach is considered of utmost interest, as is also the case with developments taking place in related areas such as knowledge representation in artificial intelligence and lexicography. These have already raised methods and tools that are not only more aligned with Web technology but also deal with matters and content that are very close to those of knowledge organization.

References


Also available online: http://lsdis.cs.uga.edu/lib/download/S98-changing.doc


*All web addresses in the bibliography are checked on 15 March 2002.*
Vanda Broughton
School of Library, Archive and Information Studies
University College London, UK

Facet Analytical Theory as a Basis for a Knowledge Organization Tool in a Subject Portal

Abstract: The paper examines the way in which classification schemes can be applied to the organization of digital resources. The case is argued for the particular suitability of schemes based on faceted principles for the organization of complex digital objects. Details are given of a co-operative project between the School of Library Archive & Information Studies, University College London, and the United Kingdom Higher Education gateways Arts and Humanities Data Service and Humbul, in which a faceted knowledge structure is being developed for the indexing and display of digital materials within a new combined humanities portal.

1. Introduction and background to the project

The work described in this paper is currently being carried out at the School of Library, Archive & Information Studies at University College London (SLAIS), under a grant from the United Kingdom Arts and Humanities Research Board. Partners in the research with the School are the Arts and Humanities Data Service (www.ahds.ac.uk) and Humbul (www.humbul.ac.uk) two large government funded subject gateways for the humanities, serving the UK higher education community. The Arts and Humanities Data Service (AHDS) comprises several distinct digital collections, including archaeology and the visual and performing arts; Humbul has a more conventional humanities content with strong collections in history, philosophy, theology, literature and the classics. The two gateways are soon to merge into a single Humanities portal.

The object of the research is to develop a subject tool for the management of the new portal with two functions envisaged. Firstly the system will be used to organize the front end of the portal, and will provide a structure for the first point of entry to the site. A directory style layout will be provided with subject headings developed from the system, together with browsable indexes derived from the inversion of the subject headings. The knowledge tool has the additional function of providing a vocabulary for use in subject metadata.

Currently the two gateways create catalogue data for all of the digital objects in their collections using the Dublin Core. For the subject field (and for organization of the display of resources) Humbul uses the Dewey Decimal classification, and AHDS Library of Congress Subject Headings. Neither of these systems have proved entirely satisfactory, and for the combined portal both organizations decided that they wanted to explore the possibilities of a new and more satisfactory indexing tool, logically structured and based on modern theories of subject organization, and designed with the indexing needs of digital materials specifically in mind.
Although it might be argued that the problem of identifying the semantic content of an information carrier does not differ with the nature of that carrier, nevertheless there are some distinctive features of digital materials that do affect the indexing process. Firstly the intellectual content of the materials can be very complex, and the level of indexing is required to be high, in terms of both the specificity and exhaustivity. The fairly broad level of subject description provided by the conventional systems currently used has created difficulties, particularly for AHDS which consists of several discrete subject oriented databases. Where there are only one or two levels of hierarchy available for index description the resulting lack of specificity often means that the item will not be retrieved by a searcher starting from the context of another discipline, since the content relevant to his discipline will not have been identified and tagged. For example, a resource dealing with nineteenth century political cartoons deposited in the Visual Arts Data Service database will not be retrieved by a searcher with an interest in nineteenth century political history if the item has only been described as ‘visual arts – cartoons – nineteenth century’. This cross discipline searching is a particular feature of AHDS which will migrate to the combined portal, where it is envisaged that the problem can only become magnified.

A second indexing problem related to the digital nature of the materials is the need to identify non-semantic properties of the objects i.e. properties of electronic format, media, and so on, since these may well be sought terms.

There is also a desire to exploit the major tool for ‘cross-referencing’ in a digital environment – hypertext. Hypertext is able to facilitate retrieval in at least two ways. Firstly within the visual display where it can be used to expose successive layers of a hierarchy (see below) maintaining the detail of the developed structure without exposing too much of it to the user at once. Secondly it supports the notion of multiple points of access to the network of resources and reinforces the regularity of the structure of the system.

During discussions with staff at SLAIS it was decided that a new indexing tool, built on faceted classification principles might meet the specific indexing needs of the situation better than the conventional systems currently being employed.

2. The nature of classificatory structures

At the end of the nineteenth century the conventional bibliographic or library classification scheme was the first means of organization of, and thus access to, information on a subject basis. Since the advent of the web we have seen a number of cases where traditional bibliographic classification schemes have been employed for the organization of web resources, usually within organized gateways or digital libraries, and usually within the academic sector. The Universal Decimal Classification is probably the most widely used, closely followed by the Dewey Decimal Classification, with Library of Congress Subject Headings and various national classification systems also represented. The recent HILT (HILT, 2001) project, which investigated means of subject access to digital materials in the UK information sector, worked with forty-two stakeholders representing libraries, archives, museums and electronic information providers; of these, the majority were using locally created systems. There is thus a proliferation of systems of subject organization on the web.
Despite this, the classification scheme is often seen as ineffective as an organizer for electronic resources (Koch, 1998). Some obvious reasons for this can be found in the lack of logical structure in older systems of classification. Traditional classification schemes are built on the basis of a tree structure, with the emphasis on downward subdivision into smaller and more specific classes; often, the only relationships acknowledged are those of super- and subordination and there is no provision for syntactic relations. As a result the classification is usually relatively broad and there may be limited facilities for combining between classes, or for expressing complex semantic content. As seen in the example given above, this can cause difficulties when searching in a multi-disciplinary environment, or when dealing with objects of a complex nature.

Classifications built on the facet analytical model work on different principles, starting from the constituent concepts of the subject, rather than from a view of knowledge in its entirety. It is important to establish what precisely we mean by facet analysis, since there are a number of interpretations of this theory, some of them very different from the original work of S. R. Ranganathan. The situation is complicated by the current interest shown by computer scientists in this area, and the frequent use of the term faceted classification to mean any system of subject organization in which analytico-synthesis is used. This notion is also often employed in the world wide web literature on facet analysis (Maple, 1995), and also in the United States usage of the term when compared with the British or European tradition. A recent commentator on types of classification scheme (Koehler, 2000) states that "Universal classification schemes .... tend to classify works by a single characteristic.", contrasting them with faceted schemes which are considered to express more than one aspect of a subject.

In this paper, and in the accompanying project, facet analysis is interpreted in a much more restricted way; it is taken to mean that rigorous process of terminological analysis whereby the vocabulary of a given subject is organized into facets and arrays, resulting in a complex knowledge structure with both semantic and syntactic relationships clearly delineated.

3. Applications of faceted classifications

Classifications of this type provide effective tools for vocabulary management, document description and retrieval. Facet analysis provides a method which is in theory appropriate for the management of terminologies and concepts in a variety of environments, although its applications have so far mainly been limited to the conventional print based document collection.

As an ordering device for initial access to resources in a managed site the faceted classification has some advantages over traditional schemes. The predictable format of the classification is usually very evident to the casual searcher, and the structure can also be used to generate subject headings and a browsable index, as alternatives to, or in support of, a directory style front page.

Although primarily used for the physical organization of print based materials the potential of facet analysis for the management of documents in an automated environment has already been recognized. More than ten years ago Godert (Godert, 1991) and Ingwersen and Wormell (Ingwersen & Wormell, 1992) testified to their performance when used in conjunction with online databases for both document description and the framing of queries. Ingwersen and Wormell
(Ingwersen & Wormell 1992, p.199) were able to state with confidence that "... the discussion demonstrates the suitability of the faceted categorization, not only for textual documents, but also with other forms of carriers of information. Faceted categorization may provide multi-dimensional and hence structured entry points to document contents, and thus give intellectual access to generated and stored knowledge." More recently the specific application of facet analysis to searching the World Wide Web has been discussed by Ellis and Vasconcelos (Ellis & Vasconcelos, 2000) who conclude that "...because ...the classification is derived inductively from the concepts or terms used in the subject field, it can alleviate some problems in searching the World Wide Web by being applied to using subject directories or search engines."

4. Facet analytical methodology

The system currently under development at UCL is built on those classificatory principles developed by the Classification Research Group in the mid twentieth century out of the original work of S. R. Ranganathan. The internal logic of the system is based on rigorous analysis of vocabulary, whereby terms are sorted into a standard set of functional categories. Within these categories a range of semantic relations are acknowledged, and problems of vocabulary control addressed. A sophisticated system syntax provides for the ordering and combination of terms both intra- and inter-category. There is consequent improved performance in the accommodation of complex subjects, the predictability of location, and in the effectiveness of retrieval. This methodology has been brought to a high level of sophistication in the second edition of the Bliss Bibliographic Classification (Mills & Broughton, 1977 - ) where it has been used to create a new general scheme of classification (albeit built on the infrastructure of the original scheme (Bliss, 1940-1953)).

Taking the BC2 methodology as a standard, the new system begins from the literary warrant of the objects to be classified. The digital objects within the humanities collections are used to establish the terminology to be worked on. A set of standard categories is used for analysis which follows the 'standard citation order'. These categories are functional in nature; they comprise thing/entity ~ kind ~ part ~ property ~ material ~ processes ~ operations ~ patient ~ product ~ agents ~ space ~ time. Although developed initially within the areas of science and technology (so that they resemble a production process), in the process of the BC2 work these standard categories have been applied successfully to the vocabulary of subject fields in almost all areas of knowledge. As one moves across the spectrum of disciplines from science to the arts it is true that the number of these categories that are used becomes smaller; notions such as materials, products and agents are less often found in the humanities. It is also true that some domain specific categories can occur. Commonly recognised examples include those of genre and form in literature and music. We expect that new categories can be determined and that the techniques of facet analysis can be applied in new ways and to new material. It is the methodology that is here being tested, and not any specific structure that has been created in the past.

Once the terms have been sorted into categories, further analysis structures each category and clusters the terms into arrays or groups of terms which share a common property or characteristic (known as the characteristic of division).
Various principles of ordering are used to determine the sequence of arrays within the category, and the collocation of terms within the array; chronological, developmental, spatial and physical contiguity, and the commonly used classificatory principles of increasing concreteness, increasing speciality, and increasing complexity. The organization of the categories or facets into a single sequence normally follows standard citation order i.e. the order given above, but inverted; this brings the more concrete facets to the end of the sequence, and facilitates the location of compounded terms.

The classificatory structure built in this way will consist of single simple terms arranged in a systematic way according to the logic of the system – a pure faceted structure in Ranganathan’s understanding. In practice we find that the terminology does not always lend itself to absolute reduction of this kind. There are constant occurrences of terms which represent compounds and combinations of all sorts, but which require to be included in the schedules (and perhaps more significantly in the index) because of their status within the literature and their existence as sought concepts. Consequently the classification consists of rather more than the ‘skeleton’ of simple concepts in mutually exclusive sets. As examples of compound concepts are used to populate the classes the structure grows in complexity. Where, as is usually the case, the compound represents the intersection of two or more categories, the system syntax (or rules of combination) control the precise location of the compound and a much expanded classification is generated by the interaction of the syntax and the conceptual structure. The basic citation order can be repeated as necessary at any point within the classification, and subjects of considerable complexity can be accommodated in a highly logical, predictable and regular fashion.

A fully developed faceted classification on the model of BC2 can by this means generate a very complex knowledge structure of n-dimensionality and great logical regularity, and with deep levels of hierarchy. The resultant structure can be utilised in a number of ways; as an ordering device, as a source of index terms and subject headings, and can also be converted to a thesaurus.

The principles of categorization implicit in facet analytical theory can be extended to organize any set of properties of objects in any domain, and of the domain itself. The theory is sufficiently well established to allow variation in the classical form, and the compiler of a faceted structure need not feel restricted to the categories and combinatorial rules of standard citation order. Viewed in this way, facet analysis should be thought of as a powerful methodological tool rather than a specific arrangement of topics in a given field. Any attributes of objects which are significant for retrieval can be incorporated into the categorical formulation; hence facet analytical theory can accommodate concepts descriptive of a digital environment and the objects in it when building a structure to function as a retrieval tool in that environment. Its capacity to cope with objects with complex combinations of attributes in a logical and predictable fashion permits the creation of structures which reflect the multidimensionality of such objects. There also appears to be potential to exploit hypertext to generate links across such structures, and to increase the points of entry to the semantic network in an innovative manner.
5. Implementation of the knowledge tool and operational testing of the system

As we have seen above, the effectiveness of browsing across a series of semantically discrete databases is important to the success of the combined portal; therefore experiments will be carried out to test its use in cross-disciplinary browsing and retrieval of digital resources within the digital collections. In a test-bed implementation for the research, AHDS and Humbul are applying the knowledge structure to the Portal’s planned metadata repository for all the digital objects in their collection. Initially a pilot project will build a structure for the discipline of history and test its effectiveness over this part of the collection. This area has been chosen because of the pervasiveness of the discipline, and its particular suitability for testing cross-disciplinary searching and retrieval. Extensible markup language (XML) will be the vehicle for implementing the knowledge structure, and the potential of this combination of the knowledge structure and the markup language seems to be very considerable.

6. Markup languages and semantic content

It is certainly true that, to date, markup languages have not been utilised to index the intellectual content of documents. Hypertext markup language is engaged only with the format and display of documents. Extensible markup language has taken things a stage further in considering document structure and provides more information to the searcher in terms of the constituent parts of the document, although from the viewpoint of traditional retrieval this leaves us only with the equivalent of the author-title catalogue organized by a descriptive cataloguing standard. The problems of intellectual content of the object have yet to be addressed.

There is currently much interest in the development of the ‘Semantic web’. In an article in Scientific American Tim Berners-Lee (Berners-Lee, 2001) looks at how the emphasis in web searching might be switched from a provenance based way of handling documents to one linked to the semantic content; in this new approach to document analysis the ontology has an essential role to play, and it is clear that Berners-Lee’s understanding of the ontology bears a close relationship to the kind of structure that we are developing at SLAIS. He says (Berners-Lee, 2001 p 34) “The most typical kind of ontology for the Web has a taxonomy and a set of inference rules. The taxonomy defines objects of classes and relations among them …..Classes, subclasses and relations among entities are a very powerful tool for Web use...Inference rules in ontologies supply further power.” He also describes (Berners-Lee, 2001 p 37) what is essentially the process of mapping and of the use of a classification with a notation to enable information exchange “An essential process is the joining together of subcultures when a wider common language is needed. … The relations allow communication and collaboration even where the commonality of concept has not yet led to a commonality of terms.”

The capacity of the classification scheme or controlled indexing language to perform this function of intermediary in the exchange of information across languages and cultures is well represented in the professional literature. We are also clear that the use of a faceted structure has advantages in terms of its efficiency as a retrieval tool, and that it has the potential to handle complex digital objects and facilitate cross-disciplinary searching. It remains to be seen whether the
implementation of a system with these properties using XML might provide us with a tool with far wider application.

Reference
Godert, Winfried (1991) Facet classification in online retrieval International Classification 18(2) 98-109
Ingwersen, Peter and Irene Wormell (1992) Ranganathan in the perspective of advanced information retrieval Libri 42 184-201
Koch, T. (1998) The role of classification schemes in Internet resource description and discovery; work package 3 of Telematics for Research project DESIRE (RE1004) www.ukoln.ac.uk/metadata/desire/classification (Accessed 29.08.01)
Mills, J. and Vanda Broughton Bliss Bibliographic Classification 2nd edition London; Butterworth and Bowker-Saur 1977
Planning Controlled Vocabularies for the UK Public Sector

Abstract: In the UK, the aim to make public sector information much more available to the citizen has led to establishment of an “e-Government Interoperability Framework” based on a set of core standards. Among the standards is a controlled vocabulary, known as the Government Category List (GCL), used to select keywords for the metadata of all electronic resources originating from central or local government. The GCL is a small and simple taxonomy, designed to facilitate high-level browsing rather than deep searching. Specialized thesauri for particular subject areas may optionally complement the GCL. To ease the indexing burden, GCL terms will often be selected by direct mapping from the specialized vocabularies.

1. Introduction

Today almost all governments are feeling both the pressure and the opportunity to make their own information freely available to the public. The opportunity extends beyond provision of information to enabling electronic delivery of services, such as the filing of tax returns or applications for passports. Furthermore, there is widespread recognition that it is not enough just to release large swathes of documentation on yet another website. Information should be organized in such a way that it will be easy for people to find.

But how to achieve this? In the UK, one thrust is to set standards for “interoperability”, basing them on existing ones that are non-proprietary and already widely used, such as HTTP, XML and PDF. Additionally a new e-Government Metadata Standard has been prepared, based on Dublin Core and developed in partnership with the Dublin Core Metadata Initiative (http://purl.oclc.org/metadata/dublin_core/) and the governments of other countries. The Office of the e-Envoy (www.e-envoy.gov.uk) is leading UK developments.

The dream is that one day information and resources will flow seamlessly within and between all public sector bodies. The citizen at just one port of call will be able to find what he wants, from anywhere within officialdom. And one of the ways he or she ought to be able to find things is by subject. The present paper is an account of the debate and decisions on what to do about access by subject. In particular, given a “Subject” element in the new metadata standard, how should it be filled? At an early stage it was felt a thesaurus should be used for selection of subject metadata, but the question then arose, “What sort of thesaurus?”

2. Thesaurus options

A huge confusion surrounds the word “thesaurus”, which some people use to mean a controlled vocabulary conforming to ISO 2788 (International Organization for Standardization, 1986), and others imagine as a sort of electronic Roget model, automatically swapping synonyms behind the scenes. For this project, six different options were considered, as well as sundry variations on them:
Models A and B above were both seen as controlled vocabularies, in that each would comprise a finite set of terms to be used for indexing as well as retrieval. The main difference between them is in size and complexity. Model A would comprise 7000 or more terms, to cover the full span of public sector interests, with inter-relationships conforming to ISO 2788. Model B would be much smaller, with perhaps 400 preferred terms and plenty of non-preferred lead-in entries. It would have limited functionality, designed for high-level browsing and navigation rather than precision searching. Its relative simplicity would make it easier for indexers to use.

Model C is an acknowledgement that many government departments and agencies already have, or are developing, their own controlled vocabularies designed for their own sectors and contexts. A thesaurus confined to one subject area should give much better results than one aspiring to cover the public sector universe. So maybe the development of specialized thesauri should be encouraged, and central government should build a “meta-thesaurus” integrating all the vocabularies into one reference tool that shows linkages between related concepts in the different vocabularies.

Model D avoids the expense of indexing or meta-tagging at source. A search thesaurus is used only at the point of search, and the whole hassle of indexing disappears. Although there is no standard for search thesauri, several examples do exist, such as the Contemporary Thesaurus of Social Science Terms and Synonyms (Knapp, 1993), a prototype proposed for a pharmaceutical firm (Lykke Nielsen, 1998) or the synonym sets built into some search engines. The function of the synonym sets is to expand a user’s query by adding synonyms, usually strung together with OR logic. (Dextre Clarke, 2000). It is possible also to adapt an ISO-2788-style thesaurus to this type of use, exploiting the equivalence relationships for search expansion.

Model E abdicates responsibility for a pan-government vocabulary, but encourages information professionals in diverse agencies to share ideas on building and implementing specialized thesauri for their own sectors. This is certainly the cheapest option when seen from the input viewpoint, but in a cross-sectoral context does little to mitigate the end-user’s expenditure of time on fruitless searching.

Finally, Model F is a more positive move towards relying on technology rather than human indexing. A taxonomy or classification scheme would be built and would be presented to the users of portals. But automatic categorization would remove the need for human indexing. Human intervention might still be needed to define and tune the categorization rules and algorithms, but this would be very much less expensive than requiring an indexer to process every document at source.
3. Workshop discussion

If any type of controlled vocabulary is to be used for assigning metadata when documents are first input, it is essential to have the cooperation of those on whom the task will fall. There is a big contrast between coordinating document input procedures in one organization, and procuring the compliance of all the bodies in the public sector! The Office of the e-Envoy therefore held a workshop of key stakeholders, with 28 government departments and agencies represented. The above models were discussed, and the implications for implementing them.

The first thing noted about the options was that they are not mutually exclusive. It would be possible, perhaps even recommendable, to opt for more than one. Some of the models could be adopted unilaterally, either sooner or later. But with Models A and B it would be important to make an early decision because they need concerted action from all parties.

Most participants in the Workshop had an information management background and a conviction that Model A was the one most capable of yielding high-quality search results. However, they were also aware that consistent indexing requires careful quality management, and is best handled by a small team of trained staff. The reality in offices today is that web “publishing” is often carried out by the authors of the original documents, or by a webmaster who is hopelessly under-resourced for the meticulous task of adding metadata. Rarely have they been trained in indexing (often called “meta-tagging” in this context) and even more rarely in how to use a thesaurus. If a significant proportion of all the resources to be searched have inconsistent subject metatags, randomly applied or omitted, then searches will yield poor quality results and Model A falls into disrepute. When the need for training searchers was added to this scenario, Model A looked even less practical.

Model B was thought more feasible. If the controlled vocabulary is presented as an expandable hierarchy with a small number of top terms, and if just 2 or 3 mouse clicks are enough to find and select the most appropriate heading, then reasonable quality can be achieved for both indexing and retrieval. Such a simple hierarchy will not of itself enable precision searching, but is good for high-level browsing.

Model C, the meta-thesaurus, is much more ambitious in terms of the building and maintenance effort. A survey of the closest central government departments revealed that 23 specialized vocabularies were already in existence, without counting all the others in local authorities and executive agencies. Specialized software would have to be commissioned to handle such a large and complex volume of data, and a team of experts would be kept busy for years.

The search thesaurus, Model D, requires less investment and could be a promising approach. However, an immediate decision was judged unnecessary because no provision needs to be made for it in the metadata standard, and no human indexing is required. The option is still open for the future.

Model E was popular with all participants. Thesaurus building is for specialists, who have much to gain from comparing notes with others in the same business. In some sectors several organizations have overlapping interests and could well share the same specialist vocabulary. The discussion forum may be physical or electronic or probably both.

Model F is much disparaged in some quarters, but there are circumstances where human indexing is simply not available and automatic categorization is the only hope. This occurs, for example, with legacy databases or when external
resources gleaned from the World Wide Web are to be integrated with internal collections. Categorization technology is improving all the time. However, the structure, format and syntax of taxonomies driving automatic categorization tend to be specific to particular proprietary software suites. The implementers of particular portals may appropriately handle development, unilaterally without need of a government standard. So the Workshop left this, too, as an open option.

4. GCL description

The clear choice to emerge from the Workshop was Model B, the high level taxonomy. It has been named the GCL (Government Category List) and now forms part of the mandatory metadata standard. Government departments, executive agencies, local authorities and other public bodies are now developing the most cost-effective way of adding metadata to all their electronic resources.

Mandatory at the input level does not mean mandatory for display purposes. Portals implementing the taxonomy can choose between many approaches, of which one of may be seen on the Quick Find page of UK Online (http://www.ukonline.gov.uk/). Other websites will no doubt customize the display for their own target audiences.

As recommended by the Workshop, the GCL aims to be small and simple enough so that a reasonable level of consistency can be expected from the hundreds of people who will be obliged to do meta-tagging. At the time of writing it has around 360 preferred terms and over 1000 lead-in terms. The preferred terms are arranged in just 12 hierarchies, so that people browsing from the top down can easily select the most appropriate heading. Figure 1 shows the top level headings, and Figure 2 shows the full extent of the hierarchy below one of the top headings. For each of these headings the full record gives any scope note, lead-in terms and cross-references. We call the GCL a taxonomy because it is a hybrid between a thesaurus and a classification scheme, not conforming to the strict conventions required for either of these, but aiming to be user-friendly in an electronic environment.
An important aim for the GCL is to serve the citizen. The choice of terminology and structure is guided neither by academic “correctness” nor by the priorities of government employees, but by where ordinary people might expect to find things. The average citizen cannot be expected to know which government department handles which transactions, and so the GCL hierarchies studiously refuse to reflect the way UK government is currently structured.

Of course it is a huge challenge to find a structure to suit the way every ordinary citizen thinks. We all know that different people think differently. Even the same person views things from different perspectives on different occasions. Therefore the GCL is polyhierarchical, allowing more than one hierarchical access route to the same subject. And copious lead-in entries and cross-references are there to support people who just cannot find what they want by browsing.

While the taxonomy does not conform to ISO2788 guidelines for term relationships, in practical terms it is maintained using a standard thesaurus management package. It is available for download from the GovTalk website (http://www.govtalk.gov.uk/) in RTF or PDF format or as a zipped set of HTML files. The HTML version will run on almost any PC and offers a very convenient way of navigating the stand-alone taxonomy.

Suggestions for more terms, synonyms and relationships are still coming in. Realistically, they will never stop coming in. Following release of Version 1.0 in January 2002, updates are expected in May and again in September 2002. Consultation procedures are in place to help us steer clear of pitfalls along the way. The GCL will have to evolve to meet changing needs and expectations.
5. Distributed model

Since the taxonomy enables browsing rather than searching, it needs to be complemented by other tools and facilities. As already noted, many government departments and agencies already possess their own controlled vocabularies designed for their own sectors and contexts. For portals and websites in a particular sector, these specialized vocabularies may provide search tools well matched to the needs. But a big challenge confronts any agency using its own vocabulary as well as the GCL.

At the time of input, hard-pressed staff are unlikely to warm to the job of indexing once with their specialized vocabulary and then again with the GCL. Therefore it is important for each public body to develop automatic mapping techniques. Typically, a mapping table is prepared showing, for each specialized term, the closest corresponding GCL heading. Then when that term is selected as a metatag, the GCL heading is automatically added too. Because the GCL is limited to high-level headings, the mapping is almost always from the particular to the general (a much more feasible proposition than the reverse process!).

Given the coexistence of specialized vocabularies and the GCL, it is interesting to compare the chosen approach with the meta-thesaurus model. To achieve the latter, it would be necessary to import all the thesauri, that is to say, all their terms and relationships, into one compilation. Considerable complexity arises when one term has different meanings in different vocabularies, or different relationships, or is non-preferred instead of preferred. With the current approach, all those differences can be tolerated, because they are not managed centrally. The sum of all the mapping tables could be likened to a distributed meta-thesaurus. But because the vocabularies and tables are managed separately, the complex
management issues never have to be addressed. The parallel is not exact, because the mapping tables are currently used only at the time of input, not at the time of searching. But in time, ingenuity could well find a way of exploiting assemblies of mapping tables as a search aid too.

6. Conclusion

We still have a long way to go before those seamless flows of information are achieved. But increasing proportions of the population are now sitting in front of a screen wired to the Internet, expecting to be able to find what they need. The GCL is just one small piece in a jigsaw that needs to be assembled for them. The implementation of standards will be vital in fitting the pieces together. Some pieces are already in place, giving service in particular areas. Others will be added progressively, and the picture that takes shape on the face of the jigsaw will evolve to meet changing needs and exploit new technology.

Two or three years may be needed before enough organizations have applied the GCL to enough resources so that we can evaluate the effects. At that stage it will be useful to revisit the debate on “what sort of thesaurus”, and see what more of the open options should be implemented.

References


Widad Mustafa el Hadi,
IDIST/CERSATES, University of Lille III, France


Abstract: The radical changes in information and communication techniques at the end of the 20th century have significantly modified the function of terminology and its applications in all forms of communication. The introduction of new mediums has deeply changed the possibilities of distribution of scientific information. What in this situation is the role of terminology and its practical applications? What is the place for multiple functions of terminology in the communication society? What is the impact of natural language (NLP) techniques used in its processing and management? In this article we will focus on the possibilities NLP techniques offer and how they can be directed towards the satisfaction of the newly expressed needs.

1. Background

The role of terms in information retrieval is essential. Several techniques and more approaches are suggested in order to acquire and filter them for information access purposes. This paper will survey the recent trends of terminology acquisition and management in relation to information retrieval, and focus more particularly on the needs and means for building ontologies.

Since we assume that information is primarily encoded as text, information retrieval (IR) can also be an application of NLP techniques: it is vital to understand the content of a document in order to decide whether it is relevant to a given information need. Some specialists point out the need for NLP technology to build the information systems of the future (Strzalkowski, 1999).

This article will briefly account for the relevance of these techniques from the point of view of automatic terminology resources acquisition processing and management and its importance in IR environments.

The task of automatic term recognition and extraction is the task/methodology that consists in the automatic recognition or extraction of the technical terms of a domain, from texts that belong to the relevant domain (Kageura et al., 2000). It is essential that we define this task since we are examining its contribution to IR. In other words it is important to distinguish this task from automatic indexing which aims at extracting (or weighing) keywords that represent documents in order to improve IR performances. Keywords are attributes of individual documents. A set of keywords attributed to a document does not obtain an independent status separate from the document (Kageura et al., 2000). In term extraction, a set of terms can constitute a terminology, related to, but not dependant on a set of documents from the domain to which the terms belong. The main goal of automatic term extraction is to retrieve the domain specific terminology and not the terms that happen to occur in a given set of documents. Whatever the target of the application/task may be, (i.e. term extraction characterizing a domain or keyword extraction representing individual documents for information retrieval purposes),
two basic approaches can be identified: the statistical one and the linguistic one. In the first case, the output of extraction should be evaluated as a set and not on the basis of individual terms/keywords describing a document. Moreover, in the case of term extraction task (the specific terminology of a domain), the set of terms should be extensive and consistent (Kagera et al., 2000) this is not the case for keyword extraction applied to IR. Even though NLP in IR is much debated (Spark-Jones, 1999) some examples show the relevance of this technology in keyword extraction for IR (Jacquemin, 1999; p. 2001). Many authors agree nevertheless on the obvious role of NLP technology in information extraction, information management and knowledge management contexts (Spark-Jones, 1999; Maybury, 2001; Bontcheva et al., 2001, among others). Others think that the success of NLP technology depends on a more radical change of focus.

The article is divided in two sections: firstly we present an overview of new trends in terminology processing and management and their impact on IR activities. Secondly, we put forward how the existing tools can be geared towards information and knowledge management applications, namely for building domain specific ontologies. Our observations and suggestions are based on an evaluation research project dealing with automated term and semantic relations extraction tools, that we conducted during the last four years (Mustafa el Hadi et al., 1998, p. 2001).


NLP techniques, corpus linguistics, artificial intelligence and the availability of electronic corpora in many languages have dramatically influenced the study of terms and their formation, thus questioning the traditional approaches. Theoreticians and practitioners believe that terminology can only be studied in discourse and that it makes more sense to accept that the term is the starting point of terminological description rather than what was traditionally called the concept. The claim is that what is named by the same term in different texts can be shown to have different referents. This position seems to fit the new framework of terminology building and acquisition tools since they are corpus-based dependant. The current requirements in this field have stimulated the needs for enriched terminological data and for their extraction from electronic corpora and ultimately had significant impact in the field terminology processing for information access and management.

The availability of NLP tools combined to the availability of electronic corpora facilitated terminology resources acquisition. This created more elaborate needs and relevant techniques. The coarse extraction techniques as (Bourigault et al., 2001) pointed out, tend to produce huge lists of undifferentiated lists of terms that are of little use for efficient document indexing and retrieval. The last evaluation campaign we conducted for evaluating terminology acquisition tools (Mustafa el Hadi et al., 2001) confirms this position. The lists of terms produced, covering two domains were of no use for manual indexing, one of the evaluation tasks. In order to be of any use this type of lists of terms must be subject to a structuring which is an important step towards exploiting extraction results.

Some efforts were undertaken to avoid the over productivity of terms. Peter (2001), suggests to select only faceted terms, i.e. terms that pertain to the domain of the query. Hongyan et al. (2001) suggest solutions to terminological disambiguation in order to avoid conflating morphologically related terms with no semantic
relationship. These two methods allowed accurate indexing and enhanced precision in document retrieval (Bourigault et al., 2001).

In the same direction Jacquemin et al. (1999); Jacquemin (2002) suggested an approach to automatic indexing over controlled vocabulary which accounts for term variation. The approach, considered by the author as a rich indexing \(^5\) combines a part of speech tagger, a generator of morphologically related forms and a shallow transformational parser. The FASTER system, is applied to French language. The evaluation of this tool on indexing task gave encouraging results (Jacquemin et al., 1999) as far as improvement of indexing through variant extraction is concerned. Moreover, these authors, point out the importance of exploiting lexical and terminological data as being the key of success in such approaches.

From a general point of view the tools we have evaluated in our previous work, especially term extraction ones, only have a limited interest for indexing tasks which is not the case with other tools (semantic relation extractors). Some of the semantic extraction tools are adapted for indexing tasks among their other applications Iota and Loria, for instance, (Mustafa el Hadi et al., 2001).

Other techniques aim at improving term and environment extraction quality. The studies in this area concern collocations and specialized verbs \(^6\) (L'Homme, 2000), relational adjectives \(^7\) (Daille, 2001).

3. The Emergence of New Techniques

Here we will briefly account for some of the innovative techniques involving terminology acquisition:

Techniques used in cross-language information retrieval (CLIR): One of the requirements is the access to multilingual data with cross-language connections between terms and their equivalents in other languages. Akiko et al. (2001) show how bilingual keywords lists frequently found in scientific journal articles can be used to build multilingual keywords clusters. Such clusters can be used for query expansion in cross-language information access and even for monolingual query expansion (Bourigault et al., 2001). Other significant contributions in this area are suggested (Bisson, 2001; Hull, 2001, Gaussier, 2001). It would be interesting in this respect to design term extraction tools for generic bi-lingual production, allowing ad hoc extractions through ad hoc interfaces (Mustafa el Hadi et al., 2001).

Techniques used for automatic text summarization: It is an emerging activity in information access in which the semantic role of terms is of vital importance (Oaks et al., 2001).

Techniques used for building ontologies, used in the emerging discipline of knowledge management \(^8\) (KM). This discipline needs a whole range of NLP and technologies to enhance IR and extraction, text summarization, information filtering. Ontologies will be examined in the following section.

3.1. Ontologies

Ontologies have emerged as one of the most popular means of modelling the knowledge of a domain. The meaning of this term varies in the literature, but as (Bontcheva et al. 2001, p. 18) pointed out: “it is minimally a hierarchical taxonomy of categories, concepts and terms of a domain. Ontologies can act as an index to the memory of an organization and facilitate semantic searches and the retrieval of
knowledge from the corporate memory as it is embodied in documents and other archives”. This is also our opinion on the subject.

For Poli (1996, p. 313), “an ontology is not a catalogue of the world, a taxonomy or a terminology. If anything, an ontology is the general framework within which catalogues, taxonomies and terminologies may be given suitable organization (...). reality is organized into diverse levels and there are sophisticated dependencies among these levels and within them. An acceptable ontology should be able to model all these relations of dependencies”.

Though they are different from concept lexicon, classification schemes and thesauri, their inter-relationship is still obvious. This is why we do think that terminology resources are the basis for this type of organization. We assume that a hybridized approach for ontology elaboration (a formal and linguistic one) is necessary (cf. relevant experiences below).

3.2. Why Build Ontologies?

Uschold, quoted by Vickery (1997) points out the need for ontologies: “Knowledge engineering suffers from lack of formal tools for understanding domains of interest. Domain conceptualization is a necessary part of knowledge acquisition for a knowledge-based system. There is need for knowledge-level analysis, a requirement for the development of knowledge-based systems (Vickery 1997: pp. 277-286)”.

Ontologies can be seen as databases with information about categories and/or concepts of a domain of knowledge as well as their relations. They are seen as “enriched” thesauri since they include definitions and semantic relations. Their role is to reduce conceptual and terminological ambiguities by identifying and properly defining a set of relevant concepts characterizing a domain. They can solve semantic ambiguity, a core problem in text understanding. In this respect, MT is increasingly seen as an important element of success in CLIR an activity which involves elements of both MT and IR. There are many other uses mentioned in the literature, but the most relevant ontology models as far as our approach and interest in this subject are concerned, are GALEN (for its multilingual terminological module), CYC (especially for the levels enabling natural language understanding and generation) and MIKROKOSMOS (for MT application). Moreover, this type of ontologies could be useful in MT environments and consequently in CLIR. This point will be examined intensively in the light of the research study we are currently conducting.

3.3. How can Terminology Resources be Directed towards Building Ontologies?

3.3.1 Automation at the Stage of Term and Semantic Relation Extraction

Most of the methods used in building ontologies are rather empirical. Two main research trends can be pointed out: the linguistic approach which is represented by the research done by Terminology and Artificial Intelligence teams (Biebow et al., 1999) and the formal one Gómez-Pérez, 1999). As far as the former approach is concerned, terminology acquisition tools (an expression which covers respectively, term extractors, classifiers and semantic relation extractors) can be used in building ontologies. Term extractors that are the most available tools
produce as output, holistic lists of terms, drawn from a representative corpus, that characterizes and describes a field of knowledge. In order to be of any use, this type of lists must be subject to a structuring which is an important step towards exploiting extraction results. Terminology resources are hence increasingly seen as structured data - i.e. as a network of terms organized by relations. Pure alphabetical lists can hardly be used except for bilingual reference tools or by specialists in the relevant fields, which makes it a rather constraining process. Structuring lists of terms by semantic relations or in classes is useful for the following applications: Index-making for on-line technical documentation; browsing, information access and retrieval, building thesauri and ontologies for information systems and recently text summarization.

The first step in building ontologies is to acquire terminological resources from corpora in different languages\. In our case, the idea is to use available terminology acquisition tools (for a complete account of these tools, cf. Mustafa \textit{et al.}, 2001). This step can be considered as an assemblage of "basic level" tools. 3.3.2. Term Extractors, Classifiers and Semantic Relation Extractors: Cooperative Integration of Tools

Many applications and extraction methods relevant to these tools have been described in the literature. But the systems tested within the framework of our evaluation campaigns are geared towards a variety of applications, ranging from rough semantic relation extraction through indexing, thesaurus construction, to knowledge-based system modeling.

The output of the systems diverge but can in some cases be complementary. In fact the results drawn from the first evaluation campaign (trial run) in 1997 (Béguin \textit{et al.} 1997) have led us at that time to consider the feasibility of trans-systemic integration for strengthening their automatic terms identification capabilities. The idea is to combine two or three different types of systems in order to specify various integrated production processes. Systems could increasingly be seen as parts of these integrated production processes.

The formal level of the ontology will be built either by integrating a tool or by conceiving from scratch an inference engine. Thus the type of ontology we are suggesting would be semi-formal (see above the definitions given by Uschold). In fact, part of the organization would be expressed in natural language.

4. Relevant Work and Similar Experiences

Using the existing term extraction and recognition tools together with semantic relation extraction tools and classifiers can be compared, though the procedure will be different, to several endeavors in the field of ontology constructions. It follows in the path of comparable work in the field of NLP applications to MT such as the SENSUS-based ontology (for more details and relevant examples see Femández López, 1999); TERMINAE (Biebow \textit{et al.}, 2000) a tool for generating ontologies from specialized corpora. It uses lexical and terminological tools to build the terminological and conceptual data enriched by definitions drawn directly from the relevant corpus. The tool used enables contextual feedback in order to pick up definitions and check their validity. The process of validation and normalization of terms and definitions are carried out by a
domain specialist. The final and formal layer is based on ALNR, logic description language as the designer of the systems mentioned Biebow et al. 2000). Other experiences are oriented towards building ontologies for knowledge management and knowledge sharing. SymOntos (Veraldi et al. 1999) a symbolic ontology management system is an example of this family of tools. The system uses OPAL (Object, Process, and Actor modeling language), a methodology for the modeling and management of a company knowledge-base. It enables the representation of the semi-formal knowledge. SymOntos gathers terms denoting concepts, definitions explaining (in natural language) the meaning of concepts and a set of relationships.

5. Future Directions

The quality and good coverage of an ontology will depend on the representativity of the corpus used in acquiring the terminological resources (terms, semantic relations and concept definitions). The first step would be the selection of a representative corpus covering a field of knowledge. The second step would be to choose among the tools we tested the most relevant ones for an integrated production of terms, semantic relations and classifications. It is important to choose tools that will allow contextual feedback in order to pick up definitions and check their validity, an essential element of the ontology. In fact, there are a variety of tools in the public as well as in the private sectors. It is possible, either to combine two or three tools or get the tested systems output (i.e. lists of terms, semantic relations and classes of terms) processed by these tools before the drawing up of the ontology. The only point remaining is the formal layer choice that is still under consideration.

Notes
1. “Ontology can be defined as a set of concepts (entities, attributes, processes), their definitions and their inter-relationships. An ontology may take a variety of forms, but will necessarily include a vocabulary of terms and some specification of their meaning (i.e. definitions). It may be highly informal, expressed in natural language, - semi-formal, expressed in restricted and structures form of natural language, - semi-formal, expressed in an artificial formally defined language, - rigorously formal with meticulously defined language terms with formal semantics and theorems”, (Uschold, 1996), quoted by Vickery (1997, p. 278).
2. This activity covers term extraction, semantic relations identification and extraction together with term classification. These distinctions go back to our early work on evaluating the relevant tools: term extractors, semantic relation extractors and classifiers (Mustafa el Hadi et al., 1998; p. 2001).
3. Most of the extracting tools consider terms as noun phrases. Systems identify terms by using frequency, distribution and categorial pattern matching. All lexical units contained in a given text are analysed and matched to patterns (typical forms of terminological units) described in rules. The linguistic approaches are heavily language-dependant, hence extension to another language involves a thorough redesigning of programs. These approaches offer a set of potential candidate terms in a pre-defined morpho-syntactic format. A statistical module is used to rank the selected terms according to their relevance to the field.
4. Two corpora served as test material Spirale, periodical (educational sciences) INRA (biotechnology of animal reproduction).
5. **Rich indexing** is opposed to **simple indexing**: the former means extraction of multi-word term occurrences and multi-word term variants; the latter means: extraction of multi-word term occurrences without considering variations (Jacquemin et al., 1999: 69).

6. **Collocations** are opposed by L’Homme (2000) to **specialized lexical combinations**. The former comprises general combinations, the latter specialized ones. They can comprise specialized verbs, specialized adjectives.

To improve term extraction quality and relevance especially in indexing, technical writing and translation task, the identification and retrieval of these elements is becoming increasingly popular in NLP applied to term and term extraction (L’Homme 2000).

7. As for relational adjectives included in noun compounds, an interesting study was carried out by Daille (2001) showing how these adjectives can be useful in several types of NLP applications. In the context of terminology resources for IR, the method suggested by (Daille, 2001) aims at grouping synonym forms referring to a unique concept. The method is used in updating thesauri.

8. Ten areas in which language technology can enable KM have been identified: input analysis, content-based information retrieval, information extraction, question answering, translation, dialogue management, user modelling, summarization, presentation / generation and awareness/collaboration (Maybury, 2001).

9. There are two possibilities: either adding a formal layer or integrating from-the-shelf tool. For lack of space we cannot develop this part.

10. Most of the tools we studied are monolingual extractors. Bilingual or multilingual resources can otherwise be acquired by alignment technology.

11. This means an agreement of the designers who participated to the AUF (Association des Universités Francophones) evaluation campaign conducted by our team at the University of Lille 3. If the tools can not be, for any reason, made available for our ontology project we have to resort to industrial products of the same type.

12. As (Chaudiron 2000), pointed out “basic tools” are a category of “low level” tools. Their role is to process a precise and limited linguistic function: lemmatisation tools, morpho-syntactic and/or semantic analysers, terminology extractors,... etc. These tools are considered as first level tools because the process they achieve is an elementary one, even if they can be very complex from the technological point of view. They create low added-value for the end user. In this sense, the output of the system is not the result of an information processing but is the result of linguistic data processing. There is no information management.

13. In some cases ontology designing will use existing ontologies manually created ones which can be enriched thanks to increasingly common term acquisition tools.

**References**


Academic Library Gateways to Online Information: A Taxonomy of Organizational Structures

Abstract: Reports a preliminary analysis of organizational schemes applied by academic libraries worldwide to arrange their electronic resources on their Web-based information gateways. The unsystematic sample consists of 41 academic libraries in 10 countries representing 4 languages, Chinese, English, German, and Spanish. The study reveals a widely accepted practice in applying 6 simplistic methods to organizing online information: by resource type, alphabetical by title, alphabetical by subject (mostly discipline and genre), by vendor/publisher, by broad classification, and random. In addition, it notes a marked difference between libraries in the English-speaking world and those in other countries in that the former present significantly more systematic characteristics.

1. Introduction

Online information has become a critical part of library services worldwide. Academic libraries in particular have steadily increased their online resources, which range from commercial databases to freely accessible sites on the World Wide Web (WWW). Interestingly, librarians appear to treat online collections somewhat differently from physical collections. Many academic libraries do not catalog online resources, or catalog only selected ones on a limited basis. Instead, they frequently provide access to online information via so-called “gateways” on their libraries’ Websites, bypassing the library catalog. The basic design of a gateway often involves the use of hyperlinks giving users direct access to individual online items.

The organizational schemes that are created to provide access to online collections in the academic library are of particular interest for a number of reasons. First, they are created by librarians, who are presumably familiar with many more possibilities for the organization of Web resources than the average person. Second, they are more apt to be extensive, and thus be more indicative of possibilities for the organization of online collections than other Websites that are created to provide gateway type access to the WWW and other online resources. Finally, these gateways are mostly local creations that do not necessarily follow widely accepted standards for representation or organization such as the International Standard for Bibliographic Description or the Dewey Decimal Classification. This lack of standardization in organizational schemes is likely to increase variations among gateways across countries. It is also possible that the variations may reflect theoretical and cultural differences, making gateways a viable source for studying the influence of cultures and institutions on information organization.
The research reported in this paper has three objectives: (1) to investigate the means by which academic libraries in a variety of different countries provide access to online information resources; (2) to construct a taxonomy of organizational structures used in library gateways to online collections; and (3) to identify cultural differences that exist among academic libraries in different countries.

The importance of the study is of two-fold. First, such an analysis will generate a broad picture of current practices by academic libraries in organizing and providing access to online resources across various countries. This knowledge will facilitate our ability to evaluate the functions and effectiveness of library gateways and to lay some of the foundations for evaluating the functions and effectiveness of directory-like Websites in general. Furthermore, findings of cultural differences across countries will contribute to the development and improvement of cross-cultural information retrieval systems.

2. Related Literature

Up until now, research on the organization of online information has not focused on the approaches taken by libraries to the organization of all online resources. Many studies deal with the designs and effectiveness of Internet search engines that, although related, are actually concerned with a specific aspect, indexing, rather than organization in general. Zeng (1997) reports a study of information organization in the WWW virtual libraries that are basically Web pages created by individuals on a voluntary basis to disseminate information around particular subject areas. The creators/hosts of those studied sites were mostly unaffiliated with libraries and resources concerned were mostly the ones free on the Internet and small in size. There have also been reports on some experiments by a small number of libraries on application of classification to the organization of Internet resources (e.g., Koch, 1997). This study will begin to fill in the gap and present some preliminary results as to how academic libraries worldwide organize a wide variety of online information in a large scale.

3. Methods

This study of international academic library gateways consists of a preliminary analysis of Websites from 41 chosen academic libraries in 10 countries: Argentina, Australia, Canada, China, Egypt (only one English site), Germany, Hong Kong (HK), Peru, South Africa, and the United Kingdom (UK). 5 institutions were chosen from each country, except 1 from Egypt, 2 from Peru, and 3 from Argentina. Hong Kong was included for comparison to China because HK was a British colony for 99 years and might have some differences from its mother country, China. Four languages are represented in the sample: Chinese, English, German, and Spanish. Several other factors also influenced the selections. Some university ranking sources were used in selecting sites in Canada, China, and UK and most top-ranked schools were picked. Most other libraries were identified through the Web site LibDex (http://www.libdex.com). The accessibility of a site was also a determining factor.

Each site was examined to ascertain the types of online information resources that appeared in the gateway, the extent to which the library actually
cataloged those online resources in addition to providing access via a Website gateway, and the organizational structures of the gateway. After the analysis was completed, characteristics of sites from different countries were further compared for similarities and differences.

As expected, most of these academic libraries provided a variety of online resources: indexing and abstracting databases, electronic journals, electronic texts and books, electronic reference documents, and Web pages freely available on the WWW. Although they categorized resources differently, these 5 were the most commonly mentioned categories. Others included government information on the Web, statistical/numerical databases, and examination materials for students. Only the above 5 were used in the following analysis. The category of e-reference was ambiguous but nevertheless popular among some of these libraries and was thus included as a separate category. Unfortunately, many libraries limited access to the catalog to their local users and as a result, it was impossible to determine whether or not they catalog e-resources.

4. A Taxonomy

Among the libraries surveyed, all (n=41) had a home page on the World Wide Web to provide gateway access to their services and resources, a precondition for being included in the sample, and all seemed to have a Web-based online public access catalog (OPAC), but some prohibited access to the OPAC by non-members. The organizational schemes applied on these gateways had striking similarities regardless of cultural and political differences in those countries. In contrast, the variations in the organizational schemes, although present, were harder to pinpoint.

Generally speaking, there were six types of arrangement used by these libraries to organize electronic resources: by resource type, alphabetical by title, alphabetical by subject and then title, by vendor/publisher, by classification, and random. Invariably, libraries first divided e-resources into several categories by resource type. Many did so by giving resource types on the library’s home page as hyperlinks and others gave these types on another centralized page for all e-resources, usually under a name like “Electronic Resources.” E-journals and databases were usually the top two choices next to the catalog. The others are explained below in some detail (see Tables 1a-1b).

<table>
<thead>
<tr>
<th>Resource</th>
<th>Database</th>
<th>E-Journal</th>
<th>E-Book/Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Arr.</td>
<td>Title*</td>
<td>Subject**</td>
<td>Vendor Title</td>
</tr>
<tr>
<td>2nd Arr.</td>
<td>Title</td>
<td>No Arr.</td>
<td>Title</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># libraries</td>
<td>32</td>
<td>22</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1a: Type of arrangement by resource
* Alphabetical by title
** Alphabetical by subject
Among the remaining five methods, the title arrangement was by far the most common. 32 of the 41 libraries organized their e-databases alphabetically by title; 25 e-journals; 10 e-books or e-texts; and 6 e-reference documents. Some included all titles into one list. The user would need to scroll down to the appropriate entry. More provided a list of the letters in the alphabet as hyperlinks on top (or on a separate page) so that the user could jump to the appropriate letter group without scrolling down. Frequently, some libraries had this title list in addition to the subject arrangement described below.

The next popular method was alphabetizing by subject and again alphabetizing by title under each subject. A couple of libraries applied subject categories without further alphabetizing by title. Subjects on these sites ranged from broad disciplines like the humanities/social sciences/sciences to fields of studies like art history, etc., to material genres like bibliographies and dictionaries. A couple even resembled the directory trees on Yahoo! and Internet search engines. Similar to the above, some libraries kept the entire list (by subject/title) on one page and others hyperlinked from each subject term to a separate location for a list of documents under that particular subject.

Another common organizational element was vendor or publisher. Possibly because vendors/publishers were few, many libraries did not apply alphabetization in this category. This arrangement seemed arbitrary and random. Some libraries also called this element by different names, for example, platforms, sources, or systems. This method was especially popular with e-journals (n=16).

Very few libraries surveyed applied general (or any other) classification to the organization of electronic resources. A South African library used the Dewey Decimal Classification to organize its e-reference documents and an Argentina library used the Library of Congress Classification for the same type of resources. In both cases, only broad classes were used without sub-divisions. The scope of their effort was very limited.

In addition, a small number of libraries listed their resources in a seemingly random fashion with no clear organizational schemes. Four out of the 5 Chinese libraries appeared to list their Chinese resources in no discernable order, at least not apparent to one of the authors who is a native Chinese speaker.

Geography was also a consideration by some libraries. However, there was little consistency in applying it. Some libraries, for example, placed a few local resources on top of a list followed by other randomly arranged entries. In a few sites, government resources were a separate category that was organized by geography. The general tendency in applying geography was to start with resources originated locally (usually in a city) and then move to areas close to this locality (e.g., a province) and then to other places.
5. Differences Across Cultures

As stated above, all 41 libraries had a Web homepage as the gateway for library services and resources. Most of them applied some sorts of schemes to the organization of their electronic resources, categorizing by resource type, alphabetizing by title and subject being the most popular three methods. There were indeed many differences in the ways they applied each method, but with few recognizable patterns.

There was one exception in terms of cultural differences. The libraries in English-speaking places demonstrated a higher degree of attention to systematic organization and had a strong tendency to apply the second organizational scheme—ordering documents first by subject and then by title (Table 2a). For example, when organizing e-databases, 4 out of 5 Australian sites, all 5 Canadian sites, 4 out of 5 HK sites, and 3 out of 5 UK sites applied the subject/title method. On the contrary, only 1 Chinese and 1 German sites did so. Also, among the 16 sites that organized e-journals this way, 11 were in English-speaking places (Table 2b). All 3 sites that did so with e-books, 15 out of 17 with e-reference materials, and 5 out of 9 with e-Web sites were English-speaking. It is unclear at this time as to why this is the case.

<table>
<thead>
<tr>
<th>Country</th>
<th># Surveyed</th>
<th># with Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Canada</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>UK</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>China</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Germany</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2a: Comparison of subject approach to databases.

<table>
<thead>
<tr>
<th>E-Journal</th>
<th>E-Book</th>
<th>E-Ref.</th>
<th>Web Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>English Sites</td>
<td>11</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>All</td>
<td>16</td>
<td>3</td>
<td>17</td>
</tr>
</tbody>
</table>

Table 2b: Number of English vs. all sites applying subject approach to e-resources.

The differences between the Chinese and HK sites were especially interesting. One distinct difference was in the application of the subject/title method to organizing databases, as indicated above. Secondly, the Chinese libraries tended to be less concerned about alphabetizing (Table 3). Two of them did not alphabetize their database titles and 1 did not alphabetize e-journals in English and other languages. Contrary to them, all HK libraries alphabetized all their e-resources in languages other than Chinese. Thirdly, the HK libraries arranged titles in Chinese by number of strokes without exceptions. Their Chinese counterparts mostly kept Chinese titles in a random order. One Chinese library, however, ordered one list in Chinese by Pin-Yin, the official Chinese romanization system. One theory that may explain these differences is that Hong Kong might have had far-reaching British
influence due to its colonization by Great Britain for 99 years. And indeed, the main language used in constructing all HK sites was English.

<table>
<thead>
<tr>
<th></th>
<th>English Databases: Alphabetical by Title</th>
<th>English E-Journals: Alphabetical by Title</th>
<th>Chinese Databases by # of Strokes</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>3</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 3: Comparison between China and Hong Kong.

Due to the unsystematic sampling, no final conclusions could be drawn in this study. The discussion in this section, in particular, was solely based on the data emerging from the small and unrepresentative sample.

6. Implications for Future Research

In sum, there were similarities and differences among the surveyed libraries in their organization of electronic resources as manifested on their Web gateways. Although inconclusive at this time, the study had two revealing findings: (1) it showed/discovered an across-the-board attitude among academic librarians, who viewed simple organizational schemes as appropriate for handling e-resources; and (2) it uncovered a tendency in the libraries of the English-speaking world to make more effort to systematically organize e-resources than their counterparts in the rest of the world. Both have implications for future research in the use of organizational structures in information systems.

With respect to the first finding, this study indicated a simplicity in the organizational schemes for electronic resources on the so-called gateway—by resource type, alphabetical by title, alphabetical by broad subject, by vendor/publisher, by broad classification, and random. In many ways, this is typical of academic librarianship. Many academic libraries shelve their print journal collections alphabetically by title, separate from the book collection. Also as common, larger universities often establish branch libraries along disciplinary lines that house disciplinary print collections—the same kind of approach as organizing e-resources by discipline. Arrangement by vendor/publisher is, on surface, unique to e-resources. But, it is somewhat similar to organization by format in a physical environment because both are dictated by access mechanism (hardware and software) necessary for viewing the contents. The reluctance among these libraries to apply classification to organizing e-resources is also seen in academic libraries with regard to the organization of non-book physical collections. Among the possible reasons are: the smaller size of a non-book collection rendering classification unnecessary, the high cost of applying classification, and the questionable suitability of any book classification scheme for organizing non-book collections.

The authors suggest that alternatives to imposing simplistic organizational approaches to the organization of physical materials on information organization in a virtual environment be considered. After all, many traditional approaches were developed as practical ways of dealing with physical documents. It has been known that every format possesses some unique characteristics with implications for
information retrieval. Future research needs to concentrate on the effectiveness of the six organizational schemes in the online and intangible environment. In addition, more experiments and studies may be conducted to investigate the usefulness and applicability of classification and other more sophisticated schemes in organizing electronic resources, as proposed recently by several information scientists (Hudon, 2000; Williamson, 1997).

The second interesting finding is the difference between libraries in the English-speaking world and those in other countries. However, since the findings are preliminary, more in-depth research is needed to confirm the difference and to identify the reasons for this difference. It is possible that the difference was caused by factors other than cultural characteristics. For example, the size of the e-collection may be a contributing factor in that the larger a collection is, the more it needs to be organized systematically and effectively. Only when we have a better understanding of a whole range of factors will we be able to appreciate the universal and unique cultural needs that are essential considerations in the design of transnational and cross-cultural information systems.

Acknowledgements: The authors wish to thank Elaine Chu, Eva Counsell, Sara Ranger, and Lorie Vik for their assistance in data collection.

Notes
1 These included Australia, Canada, Hong Kong, and the United Kingdom. South Africa was excluded because it had 11 official languages.
2 There is no alphabet in the Chinese language. The Chinese have used the number of strokes in a character to order entries.

References


Gerhard J.A. Riesthuis,
Department of Information Science, University of Amsterdam, The Netherlands

Maja Zumer,
Department of Library and Information Science and Book Studies, University of Ljubljana, Slovenia

The Functional Requirements for Bibliographic Records and Knowledge Organization

Abstract: Functional Requirements for Bibliographic Records (FRBR, 1998), the study commissioned by IFLA, brings revolutionary changes in the way we see modern computer catalogues. The catalogue is not seen as a sequence of bibliographic records and a copy of a card catalogue, but as an interconnected network of related information. Implications of the new model for the future development of catalogues are discussed. Special attention is given to access points and relationships between entities and the changes those will bring into both the formal and subject cataloguing, and authority files.

1. Introduction

The last three decades of the last century brought unprecedented changes to both the ways libraries and documentation centres operated and to their users' needs and expectations. Library automation, development of large bibliographic databases, union catalogue systems and shared cataloguing, new forms of publishing etc. were some of the revolutionary developments. At same time libraries and bibliographic agencies were faced with the need to reduce the high costs of operation, particularly for cataloguing and indexing.

In 1990 IFLA sponsored the Seminar on Bibliographic record in Stockholm, which addressed these issues. The Seminar acknowledged the need of libraries to reduce the cost of cataloguing, and saw as focus for further research: meeting user needs associated with the use of various types of materials and the broad range of eventual new requirements for bibliographic records. One of the resolutions of the Seminar was therefore to commission a study “to define the functional requirements for bibliographic records”. The terms of reference for the study stated as its purpose and scope “to delineate in clearly defined terms the functions performed by the bibliographic record”.

In 1992 the Standing Committee of the IFLA Section on Cataloguing accepted the terms of reference and appointed a study group. The final report of the study was accepted at the IFLA Conference in Copenhagen in 1997 and published the following year (FRBR, 1998, p.2-3).

The study group described as the aim of the study “to produce a framework that would provide a clear, precisely stated, and commonly shared understanding of what it is that the bibliographic record aims to provide information about, and what it is that we expect the record to achieve in terms of answering user needs.” (FRBR,
1998, p.2). The study had two objectives. “The first is to provide a clearly defined, structured framework for relating the data that are recorded in bibliographic records to the needs of the users of those records. The second objective is to recommend a basic level of functionality for records created by national bibliographic agencies.” (FRBR, 1998, p.7).

In this paper we discuss the first objective only. To fulfil this objective the working group developed an entity-relationship model of a bibliographical database. This is quite revolutionary. For the first time the problem discussed is not: how can we make a more or less adequate surrogate of the title page, augmented with a few extra access points, or how can we automate our existing cataloguing process, but how can a computer and appropriate software help to give access to information? Of course the surrogate of the title page and the existing catalogue practice were adequate to solve the access problem as long as the only bearer of information were books or articles. Findings the books or journals meant finding the information. Nowadays the connection between content and its bearer is much looser and we have difficulties to impress upon our students the idea that some information is not on the Internet, but only in old-fashioned books or journals.

2. The entity-relation model

In the entity-relationship model a set of objects of interest or entities is defined, and the relations between the entities are listed. Further the important characteristics or attributes of each entity are identified.

Three groups of entities are defined in the FRBR (1998). Group 1 entities (with no common name, but we propose the term ‘bibliographical entities’) include work, expression, manifestation, and item. These entities represent the information traditionally reflected in formal cataloguing part of bibliographic records. Group 2 entities (‘name entities’) comprise of persons and corporate bodies responsible for the intellectual or artistic content, the physical production and dissemination or the custodianship of bibliographic entities. The third group (‘subject entities’) represents the subject of works and it includes concept, object, event, and place. Also the entities of the first and second group can be subject of works.

Relationships serve as a link between entities and enable the user to navigate within the bibliographic database and beyond. Relationships can link entities of different types and instances of entities of the same type. There is e.g. a relation between a work and all the expressions derived from it, or a relationship between the author (name) and a work as examples of entities of different type. The relationship between all expressions of one work (e.g. translations) is a relationship of entities of the same type.

The study identifies four generic tasks that the users perform while searching: “to find entities that correspond to the user’s stated search criteria (i.e., to locate either a single entity or a set of entities in a file or database as the result of a search using an attribute or relationship of the entity); to identify an entity (i.e., to confirm that the entity described corresponds to the entity sought, or to distinguish between two or more entities with similar characteristics); to select an entity that is appropriate to the user’s stated search criteria (i.e., to choose an entity that meets the user’s requirement with respect to content, physical format, etc., or to reject an entity as being in appropriate to the user’s needs); and to obtain access to the entity described (i.e., to acquire an entity through purchase, loan, etc., or to access an
entity electronically through an online connection to a remote computer).” (FRBR, 1998, p.82).

3. Formal and subject cataloguing: differences and similarities

In its traditional meaning formal cataloguing has as main function to enable the access to works in a collection when formal characteristics, which are normally present in or at the manifestations (“books”) catalogued, are known. Subject cataloguing has as main function to enable to find manifestations or items of which the content (the expressions of works) fulfill certain expectations.

There are two groups of problems in these traditional views. In many cases the user wants a manifestation of an expression. If I go to a library to get a copy of the latest novel of Patricia Cornwell, it is irrelevant if it is a hardback or a paperback, but the question if it is a Slovenian or an English expression can be highly relevant. Looking for a sound recording of the Ninth Symphony of Ludwig von Beethoven many music lovers will be far more interested in a given expression than in a given manifestation. It is impossible to say beforehand if a user is searching for works, for expressions, for manifestations, or even a given item.

The second problem is that the distinction between formal and subject cataloguing is not always easy to make. The user can use the same access points to search for a known manifestation and a manifestation on a given subject. The well-known example is title words as access points. A keyword search with the name of a person in a bibliographic database gives both manifestations of works created by the person, and manifestations of works about the person.

4. Description and access points A bibliographic record consists of three parts: a description, a number of access points and administrative data like the date the record was created. Bibliographic records are built following detailed rules. An example of such a rule set is the UNIMARC Manual (UNIMARC 1994). Many (sub)fields of the records are filled with values defined in other manuals, e.g. field 200 of UNIMARC contains the title and statement of responsibility. The values of this field is regulated in a International Standard Bibliographic Description (ISBD) and cataloguing rules like the Anglo-American Cataloguing Rules, 2nd edition (AACR2). Figure 1 gives an example, taken from the Slovenian National Bibliography.
The description is contained in the fields 010, 2-, and 3-. The 5-, 6- and 7-fields contain access points. The other fields are administrative fields. Note that there is a lot of redundancy in this record: often the same data are given twice in the record, once in the description and once as access point. The name of the author is given in field 200 and again in field 700. The difference between the two mentions is the form of the name: S. Stanic in 200 and Stane Stani in field 700. Only the form given in field 700 can be found in the index of the Slovenian National Bibliography as Stani, Stane. A search with Stanic gives no result. In the Dutch National Bibliography the book is found with Stanic but not with Stani! Of course it should be found under both forms of the names: the correct Slovenian form, and the form the Dutch publisher put on the title page.

To achieve this goal the different forms of the name should be connected with each other. There are two possibilities: put the different forms in the bibliographic record (e.g. in the 9-fields of UNIMARC), or use authority files. The first possibility would mean that for each possible form of the name a new field in the bibliographic record should be filled. With authority files it would be possible to have in the bibliographic record only the form as is given on the title page. In one record of the authority file all different forms of the name of the same person are collected. Searching with a name of a person starts in the authority file to find all forms of the name, which then are used as search terms for the bibliographic
records (as long as the user does not stipulate that she only wants to search with the form given by her). In this view it is not necessary to declare one of the forms of a name as the uniform form. The uniform heading was used in card and list catalogues to get descriptions of authors and contributors together at one place in the alphabetical order (Verona, 1963). And the alphabetical order is a necessary evil in card and list catalogues to make searching possible. The FRBR don’t require uniform headings, but it is difficult to see how the model can be adopted without authority files, at least for names¹.

Titles are a different matter. The identify task mentioned in the FRBR makes it almost compulsory to mentioned the original title or other alternatives titles in the description. This extra titles in the descriptions can be indexed and used for searching in the same way as the main title, which is given only once in the bibliographic record.

5. Access points to more characteristics and more access points per characteristics

The FRBR gives for each entity a list of characteristics that should be included in the relevant records. Some of them are only relevant for the identify, select or obtain tasks, but most characteristics are for at least some users very usable as access points. A few examples: a teacher of Dutch at the University of Bucharest can be interested if and which books in Dutch are present in the collection of the University Library. Only if he can search with language as access point his curiosity can be satisfied. For others it could be interesting to be able to search for books published in given cities, or printed before a given year. Given access to more different characteristics is expensive when one has to make an extra card catalogue or printed index, e.g. for printers or places of publications. For the present computers and database programs the extra costs can be neglected, and in many catalogues the data are present. Often however they are not available for searching or even viewing.

Another point that is a typical relict of the card and list catalogues is the restriction of the number of access points for one characteristic. When the number of authors of a work is more then three often it can only be found with the name of the first author (Braun, 1969). The AACR2 states that when there four or more persons or bodies are involved in a particular instance, an access point should be made only for the one named first, although it also said that if an extra entry is required in the context of a given catalogue, it should be made. This “rule of three” is discussed at the moment by the Joint Steering Committee for Revision of Anglo-American Cataloguing Rules². It should remembered that the “rule of three” is not in accordance with the FRBR’s statement that the catalogue should assist the user to do at least the following: “Find all manifestations embodying: the works for which a given person or corporate body is responsible ...”, nor was it ever in accordance with Cutter’s often cited requirement that the objective a catalogue is to enable a per son to find a book of which either the author, the title or subject is known (Cutter, 1876).
6. Challenges and conclusions

For the future we need systems, which will “meet user needs associated with the use of various types of materials and the broad range of eventual new requirements for bibliographic records”. There are several questions that have to be answered before such systems can be answered. We mention only a few:

1. What to do with the distinction between the content, the information and its carrier?
2. What is the basis of the cataloguing: the work, the expression, or the manifestation?
3. Connected with these two questions is the question how bibliographic databases should be built and, particularly, presented to the user.

The first question is more complex that it looks at the first sight. In most cataloguing rules there are separate rules for different types of what is called bibliographic forms. The IFLA published several International Standards for Bibliographic Description (ISBDs). They exist for Monographs, Serials, Non-Book Materials, Printed Music, Electronic Resources, Cartographic Materials, and for Antiquarian Books. The IFLA also published an ISBD(G) meant to unify the more specialized ones, and Guidelines for the cataloguing of entities that are parts of greater entities, like journal articles. The AACR2 has a chapter “General Rules for Description” followed by 12 chapters giving instructions for the description of more specialized bibliographic forms, like Book, pamphlets and printed sheets, Music, Sound recordings, Computer files, Manuscripts and Microforms. These bibliographic forms are a strange mixture of content, physical carrier, way of publishing, etc. This gives more and more problems in a time when mixture of these forms are more and more common. There have been pleas for rules, that are not organized according to bibliographic forms, but based on the different fields recognized in ISBD(G), like title, edition, imprint and so on (Delsey, 1998). In 2001 Zlata Dimec and her collaborators published Slovenian rules based on this principle (PREKAT, 2001). In our view this is the only practicable way, especially when a strict separation of data of work and expression (both content) and manifestation and item (both carrier) can be established.

The second question is and will be the subject of much discussion. A simple answer does not exist. The existing cataloguing is based on the idea that the manifestation is described and access points are chosen for the work. This principle works all right as long as the manifestation contains one work only and the title of manifestation and work is the same. When the manifestation contains more works or title of manifestation and work are not the same the difficulties can start. In the case of a translation what should be the title access point? Should it be the title of the original work, or the title of the translation (an expression) or both? The principle says the title of the work, many rules say the title of the manifestation and if wanted add an extra access point for the title of the work. In practice often both titles can be used in searching. In most instances the title of the manifestation is the only title access point in the case of more works in one manifestation, especially if there are more then three works involved.

It is possible that a user is looking for a particular expression of a given work, in a particular format, e.g. the Violin Concerto in A minor, BWV 1041 of Johann Sebastian Bach, played by Emmy Verhey in the Beurs van Berlage, Amsterdam, in 1992, and that expression on an audio-CD. One manifestation of this
concerto can be found on a CD called Violon Concertos. How can it be found? And the user is content with any CD on which the Concerto is placed. In this case an expression-oriented cataloguing will work much better than a work or manifestation oriented cataloguing! At least composer, title (in which form?) and performers should be access points. There is a change that in many existing catalogues of CDs the answer to the question is very difficult to find or not of all. Some existing bibliographic databases offer a keyword search, and then searching with BWV 1041 can help, if all the works on a CD are mentioned in an annotation.

This brings us to the third question. Future bibliographic databases should give access to all for the finding task relevant characteristics of the bibliographic entities work, expressions, manifestations, and items, with access to authority files for the name and subject entities. These authority files should enable the user to find the correct form of access points especially in difficult cases as notations of classifications and name-title access points when there are name or title variants. In this situation the bibliographic database consists of a network of connected records, e.g. a record for a work is connected with a number of records for expressions of that work and with records for manifestations in which one or more of the expressions is present. Each of these records can be connected with one or more records from authority files, which in return each can be connected with several records for bibliographic entities. In this way navigation from record to record is possible, from an authority record to records for bibliographic entities or to other authority file records.

For this network it is necessary that also for authority file records entity-relationship models will be developed. An IFLA working group is engaged in this work (Bourdon, 2001). Another task that is still waiting completion is the development of guidelines for display formats based on the networks described. Another IFLA working group is working on this.

However we can conclude that the FRBR will have a great impact on future knowledge organization. Maybe it is not going as fast as many may want, but there is a lot of discussion going. We are learning how computers can help to find information more easily and along paths Cutter, and even the participants of the Conference on Cataloguing in Paris (1963) could only dream of, or may be not even that.

Notes
1. In our view uniform headings are also not necessary for displays in cases where a search gives a great number of hits, like a search for Bibles or all editions and versions of Shakespeare's Hamlet in the Library of Congress catalogue. Reformulating a question into e.g. Bibles AND Slovenian is faster than scroll through the surrogate card catalogue on the screen till Bible.Slovenian is reached. It is ironic to read in the FRBR that parallel titles should not be mandatory in national bibliographies any more, what is not helping those whose mother tongue is a "small" language. Abandoning the superfluous uniform (and main?) heading brings far more gain in labour productivity.

References
Bourdon, F. (2001). Functional Requirements and Numbering of Authority Records (FRANAR): to what extent authority control can be supported by technical means.


Snunit Shoham and Rochelle Kedar  
Department of Information Science, Bar-Ilan University, Israel

The Subject Cataloging of Monographs with the Use of a Thesaurus

Abstract: This paper presents the findings of a study of indexing procedure with the use of a thesaurus for post-coordination. In the first phase of the study, the indexing records of 50 books, prepared by a central cataloging service (the Israeli Center for Libraries), were compared with the indexing records for these books prepared by three independent indexers. In the second phase, indexing records for three books prepared by 51 librarians were studied. In both phases, indexing records were analyzed for mistakes and possible reasons for these mistakes are offered.

1. The Literature

The concept and tools for the subject cataloging of monographs in library collections and for the indexing of articles for retrieval from databases have developed, over time, as two separate methods for information retrieval. Originally, theorists viewed the subject cataloging of monographs as an aspect of classification. Consequently, classification systems developed in the late nineteenth and early twentieth centuries offered a dual solution: a physical, subject-oriented arrangement of materials on the shelf and the construction of a classified catalog for information access. While the Dewey Decimal Classification system employed the same approach for both physical and intellectual access, the Library of Congress developed related but separate methods, i.e., Library of Congress Classification (LCC) and Library of Congress Subject Headings (LCSH). All of these methods are based on the concept of pre-coordination of subjects, which was the practical solution for subject access before the computerization of library catalog.

Indexes for journal literature of entire fields of knowledge first appeared in the nineteenth century, employing pre-coordinated classification schemes and taxonomies. With the introduction of the computerization of large databases of journal articles, new methods were sought to improve the efficiency of retrieval. Eventually, with the addition of Boolean searching capabilities, many databases employed hierarchical thesauri for the indexing of journal articles and the construction of post-coordinated queries. These developments released both the indexer and the searcher from the burden of correctly constructing pre-coordinated subject strings.

In recent years, research in the construction and use of library catalogs has shown that both librarians and library users encounter difficulties in the use of pre-coordinated subject systems. Most of the research examined the Library of Congress Subject Headings (LCSH). Franz and others' (1994) study among 63 end users reveals that the users only partially understood the meanings of the subject headings. Drabenstott (1999) identified a problem of misunderstanding of the subject headings by library users, children and adults, and by reference librarians. In a
study among 144 children and 144 adults who were recruited in three public libraries in Michigan, it was found that adults understood subject headings better than children; however, both of them assigned correct meaning to less than half of the subject headings they examined (Drabenstott, Simcox, and Fenton, 1999).

Other studies reported on misunderstanding and mistakes in the use of the subject headings by librarians. Svenonius and MaGarry (1993) found that among the 200 Subject Headings given by librarians (who were not working with LC) to a sample of 100 books, 25% were incorrect. The most common mistakes were "too broad descriptor" (about half of the cases), "off topic" (about one third) and "too specific" (only few). In another study, 46% of the meanings given by 137 reference librarians and 135 technical service librarians to subdivided subject headings were incorrect. Librarians did best when they assigned meanings to subject headings alone and did worst when they assigned meaning to subject headings in bibliographic records (Drabenstott, Simcox, and Williams, 1999).

Research has shown that many of the difficulties in using the LCSH are related to its being a pre-coordinated system, and therefore constructed from a list (sometimes long) of sub-subjects. It is hard for users to understand the logic behind the string construction. The question arises whether the users of the LCSH would show better or worse results if they were to use a post-coordinated controlled vocabulary. The present study attempts to explore an alternative method for subject cataloging of monographs - a post-coordinated system, consisting of key words (thesaurus).

In 1994 the Israeli Center for Libraries (ICL), which provides central cataloging to all public and high school libraries in Israel, decided to develop a thesaurus based on the thesaurus used by The Index to Hebrew Periodicals (IHP) produced by Haifa University, with various adaptations and changes in accordance with the above factors. The Thesaurus of Indexing Terms (Hebrew, 1996), produced by the ICL, is a controlled, hierarchy vocabulary of descriptors intended for post-coordination. Containing over 12,000 descriptors and over 5,500 non-descriptors, it has been in use in Israeli public and school libraries since 1996. Until 1996 public and school librarians did not index/subject catalog their collections at all, relying on classified catalogs.

2. Research Design

Research into the use of the Thesaurus of Indexing Terms has been carried out on two fronts:

I. A sample of 50 non-fiction monographs, published in Hebrew in 1997 and cataloged by the ICL, with proportionate selections from all Dewey classes. In 1997 the ICL cataloged 2678 items, of which 1050 non-fiction titles were subject cataloged with the use of the Thesaurus of Indexing Terms (1996). The distribution of the 1050 non-fiction titles among the Dewey classes was examined, in order to determine the relative size of the sample from each Dewey class. Table 1 presents the distribution of the research sample. Within each Dewey class the titles were chosen randomly.
<table>
<thead>
<tr>
<th>Dewey Class</th>
<th>No. of titles in ICL catalog for 1997</th>
<th>Sample size 50 * (no. of titles in class/1050)</th>
</tr>
</thead>
<tbody>
<tr>
<td>000 Generalities</td>
<td>91 titles</td>
<td>4 titles</td>
</tr>
<tr>
<td>100 Philosophy</td>
<td>139 titles</td>
<td>7 titles</td>
</tr>
<tr>
<td>200 Religion</td>
<td>67 titles</td>
<td>3 titles</td>
</tr>
<tr>
<td>300 Social Sciences</td>
<td>190 titles</td>
<td>9 titles</td>
</tr>
<tr>
<td>400 Languages</td>
<td>24 titles</td>
<td>1 title</td>
</tr>
<tr>
<td>500 Natural Sciences and Math</td>
<td>43 titles</td>
<td>2 titles</td>
</tr>
<tr>
<td>600 Technology</td>
<td>160 titles</td>
<td>8 titles</td>
</tr>
<tr>
<td>700 Arts</td>
<td>46 titles</td>
<td>2 titles</td>
</tr>
<tr>
<td>800 Literature</td>
<td>51 titles</td>
<td>3 titles</td>
</tr>
<tr>
<td>900 History</td>
<td>170 titles</td>
<td>8 titles</td>
</tr>
<tr>
<td>(9) Israel *</td>
<td>69 titles</td>
<td>3 titles</td>
</tr>
</tbody>
</table>

Table 1: Distribution of the research sample

A panel of three experienced indexers indexed these books, in order to establish a single "correct" indexing record (CI) for each of the monographs. These indexers had several years of experience in article indexing and/or subject cataloging of books, and all had experience as teachers of indexing. These indexing records (CI) were compared with the indexing records for these 50 books prepared by the ICL.

II. A survey of 51 librarians working in 48 different libraries (36 public libraries, 8 school libraries, 4 colleges) as to their use of the Thesaurus of Indexing Terms. The questionnaire posed questions as to the assignment of indexing, and the performance of searches using indexing descriptors. In addition, they were given three specific items to index. These items were chosen from the original group of 50 books, with specific instructions that the librarians index these items without reference to the ICL indexing records for these books.

3. Findings

The panel of experienced indexers (CI) assigned 200 descriptors to the fifty books, averaging 4.00 descriptors per book (sd. 1.99), while the ICL assigned 222 descriptors, averaging 4.44 descriptors per book (sd. 2.07). The number of descriptors assigned to each item was examined in relation to its assigned Dewey classification and in relation to the length of the item. When examining the number of descriptors in relation to the subject area (Dewey class), it was noted that in the subject areas of religion (DC 200) and social sciences (DC 300) both the ICL and the panel of experienced indexers exceeded their averages. In the areas of language (DC 400), natural sciences and mathematics (DC 500) and technology/applied sciences (DC 600) both groups assigned considerably less descriptors than their averages (see Table 2).

The influence of the length of the item as a factor determining the number of descriptors assigned was examined. It was generally found that as the extent of the items indexed increased so did the average number of descriptors assigned, though
for extremely long books the average number of descriptors decreased (see Table 3).

The ICL assigned a total of 222 descriptors, which contained 84 mistakes. More than half (48) of the mistakes were the assignment of a descriptor for a topic for which there is less than 20% literary warrant. The second most common mistake (21.43%) was the application of too broad a descriptor for a topic appearing in the item (see Table 3). In absolute numbers, most of the mistakes occurred in the subject area of Israel (9), while the least number of mistakes occurred in the area of literature (DC 800) (see Table 4). In relation to the length of the item we found a pattern similar to the relation between the number of descriptors and the length of the item. The number of mistakes consistently rose with the length of the book, up to 500 pages, then dropped for books over 500 pages.

<table>
<thead>
<tr>
<th>Dewey Class</th>
<th>No. of books indexed</th>
<th>CI Average</th>
<th>SD</th>
<th>ICL Average</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>000 Generalities</td>
<td>4</td>
<td>2.50</td>
<td>1.00</td>
<td>3.75</td>
<td>2.25</td>
</tr>
<tr>
<td>100 Philosophy</td>
<td>7</td>
<td>4.00</td>
<td>1.71</td>
<td>4.43</td>
<td>2.04</td>
</tr>
<tr>
<td>200 Religion</td>
<td>3</td>
<td>4.33</td>
<td>0.44</td>
<td>5.33</td>
<td>0.51</td>
</tr>
<tr>
<td>300 Social Sciences</td>
<td>9</td>
<td>5.22</td>
<td>2.30</td>
<td>5.22</td>
<td>1.14</td>
</tr>
<tr>
<td>400 Languages</td>
<td>1</td>
<td>1.00</td>
<td>0.00</td>
<td>3.00</td>
<td>0.00</td>
</tr>
<tr>
<td>500 Natural Sciences and Math</td>
<td>2</td>
<td>3.00</td>
<td>1.00</td>
<td>4.00</td>
<td>0.00</td>
</tr>
<tr>
<td>600 Technology</td>
<td>8</td>
<td>2.50</td>
<td>0.75</td>
<td>2.88</td>
<td>1.38</td>
</tr>
<tr>
<td>700 Arts</td>
<td>2</td>
<td>4.50</td>
<td>1.50</td>
<td>4.50</td>
<td>1.50</td>
</tr>
<tr>
<td>800 Literature</td>
<td>3</td>
<td>6.00</td>
<td>1.33</td>
<td>4.00</td>
<td>0.67</td>
</tr>
<tr>
<td>900 History</td>
<td>8</td>
<td>3.63</td>
<td>0.63</td>
<td>5.13</td>
<td>1.75</td>
</tr>
<tr>
<td>(9) Israel</td>
<td>3</td>
<td>6.00</td>
<td>1.33</td>
<td>5.67</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Table 2: Average number of descriptors assigned, according to Dewey Classes

The mistaken assignment of a descriptor for a topic, which did not appear in the item (no literary warrant), was less than 10% of the mistakes in the ICL records. Even this small percentage of total number of mistakes (8 out of 84 mistakes) seems surprising for the ICL. The fault may be attributed to the sparseness of scope notes in the *Thesaurus of Indexing Terms*, which would explain less known descriptors or descriptors whose geographic or chronological scope may be ambiguous. The fairly high number of mistakes in the use of too broad a descriptor (21.43%) is comparable to the findings of Svenonious and McGarry (1993). While these researchers speculated as to possible reasons for this phenomenon (e.g., indexing policy, lack of knowledge on the part of the indexer), in the case of the present study we believe that this finding may be attributed to the lack of scope notes whose purpose is to define the scope of each descriptor.
<table>
<thead>
<tr>
<th>Type of Mistakes</th>
<th>No. of mistakes</th>
<th>percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Too broad a descriptor</td>
<td>18</td>
<td>21.43%</td>
</tr>
<tr>
<td>2. Broad + narrow descriptor</td>
<td>9</td>
<td>10.71%</td>
</tr>
<tr>
<td>3. No literary warrant</td>
<td>8</td>
<td>9.52%</td>
</tr>
<tr>
<td>4. Less than 20% literary warrant</td>
<td>48</td>
<td>57.14%</td>
</tr>
<tr>
<td>5. Assignment of non-descriptor</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>6. non-descriptor + reference</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>7. Spelling</td>
<td>1</td>
<td>1.19%</td>
</tr>
<tr>
<td>8. Descriptor non-existent in thesaurus</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>9. Assigning of same descriptor twice</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>Total</td>
<td>84</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

Table 3: Distribution of mistakes by ICL

The findings of the survey conducted among the 51 librarians was compared to the performance of the ICL for only three books. The amount of mistakes and their distribution were found to be significantly different. The three books that were chosen were from Dewey (DC) subject areas 028, 510, and 778. For these three books the librarians assigned a total of 585 descriptors with an average of 3.82 per book, whereas the ICL assigned an average of 3.66 descriptors per book. While the ICL’s average of mistakes for these three books was 1.33, the librarians averaged 1.94 mistakes for book, 49.62% more mistakes than the ICL. The distribution of the mistakes can be seen in Table 5. The most frequent mistakes were the assignment of a descriptor for a subject for which there is no literary warrant in the book (28.24%), less than 20% literary warrant (18.94%), and the assignment of a broad descriptor and its narrow descriptor to the same book (18.94%). As all three books were of similar length (from 133 to 171 pages) no conclusion could be drawn as to the impact of the length of the book on indexing performance. In relation to the impact of the subject matter of the books on indexing performance, the librarians best performance was for the book from DC 510, while the most mistakes were in DC 778. The ICL performed best in DC 028, while the most mistakes were also in DC 778.

4. Discussion

A revealing finding of this study is that the mistake of the assignment of a descriptor for a topic with less than 20% literary warrant in the item ranked first for the ICL. For the librarians in the field this was also a prominent mistake (almost 19% of the mistakes). A possible explanation of this phenomenon may be found in the publishing patterns of non-fiction works in Israel and in the resulting character of library collections. Approximately 1000 to 1500 Hebrew language, non-fiction works are published per year in Israel. Israeli public libraries serve a public that overwhelmingly looks for materials in Hebrew.
Therefore, public library non-fiction collections are overwhelmingly in Hebrew, as opposed to academic library collections in Israel, which are primarily in English. Since the supply of Hebrew language, non-fiction books is limited, indexers of public library collections feel that they must index topics that are less than 20% of the item for they may have very little on these topics in their collections. ICL indexers, while aware of the general 20% rule, had not finalized their own indexing policy by 1997, nor has the ICL published a formal, written indexing policy as of yet. This would also compensate for the relatively high rate of omissions of descriptors. It is not that the indexers of the ICL nor the librarians in the field omitted the indexing of so many topics: rather, that in numerous cases they consciously chose to index these topics with narrower and more specific descriptors.
than one general descriptor. The most important findings in relation to the librarians is the surprising amount of mistakes and their distribution. The most frequent mistakes were assigning of a descriptor for which there is no literary warrant and the assignment of both a broad descriptor and its narrow descriptors to the same book. This finding seems to point to a severe lack of knowledge of indexing procedure. The assignment of a descriptor for which there is no literary warrant can only be partially attributed to the misunderstanding of descriptors.

<table>
<thead>
<tr>
<th>DC 0.28 Book 1</th>
<th>DC 510 Book 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ICL libraries</td>
</tr>
<tr>
<td>1. Too broad a descriptor</td>
<td>16</td>
</tr>
<tr>
<td>2. Broad + narrow descriptor</td>
<td>26</td>
</tr>
<tr>
<td>3. No literary warrant</td>
<td>11</td>
</tr>
<tr>
<td>4. Less than 20% literary warrant</td>
<td>7</td>
</tr>
<tr>
<td>5. Assignment of non-descriptor</td>
<td>6</td>
</tr>
<tr>
<td>6. non-descriptor + reference</td>
<td>7</td>
</tr>
<tr>
<td>7. Spelling</td>
<td>1</td>
</tr>
<tr>
<td>8. Descriptor non-existent in thesaurus</td>
<td>7</td>
</tr>
<tr>
<td>9. Assigning of same descriptor twice</td>
<td>0</td>
</tr>
<tr>
<td>Total no. of mistakes</td>
<td>0</td>
</tr>
<tr>
<td>Total no. of descriptor assigned</td>
<td>137</td>
</tr>
<tr>
<td>Average no. of mistakes per librarian</td>
<td>1.59</td>
</tr>
<tr>
<td>Average no. of descriptors per librarian</td>
<td>2.68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DC 778 Book 3</th>
<th>total</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICL libraries</td>
<td>libraries</td>
<td>libraries%</td>
</tr>
<tr>
<td>1. Too broad a descriptor</td>
<td>1</td>
<td>23</td>
</tr>
<tr>
<td>2. Broad + narrow descriptor</td>
<td>27</td>
<td>17.88%</td>
</tr>
<tr>
<td>3. No literary warrant</td>
<td>1</td>
<td>57</td>
</tr>
<tr>
<td>4. Less than 20% literary warrant</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>5. Assignment of non-descriptor</td>
<td>7</td>
<td>4.64%</td>
</tr>
<tr>
<td>6. non-descriptor + reference</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>7. Spelling</td>
<td>11</td>
<td>7.28%</td>
</tr>
<tr>
<td>8. Descriptor non-existent in thesaurus</td>
<td>11</td>
<td>7.28%</td>
</tr>
<tr>
<td>9. Assigning of same descriptor twice</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>Total no. of mistakes</td>
<td>3</td>
<td>151</td>
</tr>
<tr>
<td>Total no. of descriptor assigned</td>
<td>271</td>
<td>11</td>
</tr>
<tr>
<td>Average no. of mistakes per librarian</td>
<td>2.96</td>
<td></td>
</tr>
<tr>
<td>Average no. of descriptors per librarian</td>
<td>5.31</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Distribution of Mistakes for Librarians and ICL

5. Conclusion
In relation to the modification of the indexing tool, it became apparent that the *Thesaurus of Indexing Terms* must undergo revision, especially in the area of
the addition of more scope notes than are included at present. This would help all librarians to better understand individual descriptors and would make the task of the local cataloger/indexer much easier. However, our findings seem to detect that this new tool has not yet been fully comprehended or assimilated by librarians in the field, nor have they been prepared enough for this new task of indexing, which commenced in 1996. It seems that groundwork for the move from the classified catalog to the addition of indexing records was not fully completed, and more must be done to train librarians in indexing and the use of this new tool.

References
Application of the Cantor Set Theory in Making Decisions about the Collections Development

Abstract: The procedures by which library collections can be evaluated are quite diverse, and some are rather complex. The Cantor Set Theory is applied with a structuralist approach as a methodological aid to decision-making about the collections development. The methodology suggested here makes use of local holdings information based on an evaluative study of the Spanish university library collections.

1. Introduction

In view of the advances in technology and communication, it is increasingly important to think in global terms about the organization of knowledge while acting locally in collections development. Libraries today are inter-linked, in a variety of ways, on a huge chain of access. Many of our libraries, however, face budgetary limitations. Therefore, when making decisions about the collections development, information service professionals should rely on systems that can represent knowledge and help evaluate the utility of resources.

Knowledge representation can be described as the group of processes of notational or conceptual symbolization of human knowledge in the context of any thematic area. The representation of knowledge in library collections includes classification, indexing, and all those linguistic and informational operations involved in the symbolic transfer of knowledge (Barité, 1997).

The mathematical theory of sets put forth by Cantor can prove very useful in providing a graphic representation of knowledge to be used as an aid in the processes tied to collection development. The set theory describes collections of cases or objects that constitute entities per se (sets). The set is defined, then, as an entity containing other entities, a definition well-suited to library collections. Yet the set theory is not limited to describing the relationships of the sets with their elements, but also the relationships among the elements and subsets themselves. In dealing with the notion of set, we must point out two basic types of relations: those of inclusion or membership (of an element in a set, or of a subset in a greater set) and those of intersection of single elements that belong to different sets (Dauben, 1990).

The structuralist approaches entails interpreting the behavior and the properties of thematic areas. Behavior is understood in terms of temporal determinants, whereas the properties are the organizational principles of classification and order. The structuralist framework serves to identify a pattern of
relationships, as it is assumed that the efficient and effective the collection development depends on the identification of an underlying structure, which comprises the coincidental relations between demand, knowledge and the patterns of publication (Baughman, 1977). Once this structure is uncovered, a collection development policy can quickly be formulated.

To facilitate the comprehension of the mathematical model, we represent the behavior and the properties of a specific thematic area, in this case LIS, as described in an evaluative study of the library collections of the Universities of Salamanca and Granada, and Carlos III University of Madrid (Pérez López, 2002). The references used in the departmental scientific production and the current state of the collections were analyzed in terms of accessibility, localization and availability, and suitability of the collections insofar as subject headings, document type and language.

2. Application of Cantor’s Mathematical Theory to a Thematic Area

We shall define a universe set: Ω = {thematic area of Library and Information Science in Spanish university libraries}.

As subsets we have:

β = {Scientific production} which, in turn, contains the subsets Carlos III {subset X}, Salamanca {subset Y} and Granada {subset Z}.

Subset Δ is defined as Δ = {the collections of the three university libraries studied}, which comprises subsets Carlos III {subset A}, Salamanca {subset B} and Granada {subset C}. From the standpoint of collection development, each one of these subsets will contain other subsets that coincide with the variables and indicators evaluated. All the sets and subsets defined can present intersections. They are not disjunctive.

2.1 Representation by Extension and Comprehension

Once established which sets are to be represented, it is necessary to consider a form of notation and representation that will adequately identify the component elements. The most commonly used methods for this purpose are comprehension, when the elements are characterized by a certain property P (e.g. having been cited more than three times) cited more; and notation by extension when each one of the elements that belongs to the set is indicated. This is especially appropriate for the study of collections as it allows us to take note of the elements title by title.

We may define subset M by comprehension as M = {periodical publications referenced more than three times by the Departments of Library and Information Science of the Universities studied}.

Given our results, we may define subset M by extension as M = {Journal of the American Society for Information Science (JASIS); Scientometrics; Information Processing and Management; Information...}.

2.2 Relationships of Membership and Inclusion

Having defined the elements of the universal set {X}, {Y}, {Z}, {A}, {B}, {C}, the following relationship of membership is generated: {X} ∈ Ω, {Y} ∈ Ω, ...

Given the set Ω = {Thematic area of Library and Information Science in Spanish University Libraries}, as subsets we have, among others, β = {Scientific
183

production} which, in turn, contains the subsets Carlos III {subset X}, Salamanca {subset Y} and Granada {subset Z}.

Subset Δ is defined as Δ = {the collections of the three university libraries studied}, which comprises subsets Carlos III {subset A}, Salamanca {subset B} and Granada {subset C}.

We have: Ω ⊆ (A ∪ B ∪ C) U (X ∪ Y U Z).... Taking into account the causal relations and interrelations of inclusion, we are able to represent the extension and scope of the thematic area studied.

Fig. 1: The property of inclusion allows us to represent each and every one of the elements that make up the area of knowledge studied and the relationships of membership among them, as well as the degree of pertinence of the collection.

2.3. Intersection

In the cases of the Universities of Granada, Salamanca and Carlos III of Madrid, we have that:

A = {Collection of the Department of Library and Information Science of the University Carlos III of Madrid};

B = {Collection of the Department of Library and Information Science of the University of Salamanca};

And C = {Collection of the Department of Library and Information Science of the University of Granada}.

The intersection of A, B and C is expressed as A ∩ B ∩ C, and indicates the set comprising all the referenced works in the possession of the three university libraries studied, which we will denote as the basic core of the collections. This implies that those titles found in all three collections are equal elements: x ∈ A; x ∈ B; x ∈ C.
Segment $e = A \cap B$ represents the holdings information and the lines of research commons to Carlos III and Salamanca; $f = B \cap C$ those common to Salamanca and Granada; and $d = A \cap C$ those common to Carlos III and Granada. The non-intersecting segments represent particular lines of research or interdisciplinary research connected with other thematic areas.

3. Conclusions

Cantor's theory proves to be a very useful mathematical tool for producing graphic representations of the current state of university library collections in a specific thematic area. The quality and limitations of end results will depend on the number of sets and subsets, the P properties applied, and on the indicators used, such as accessibility, organization, localization and availability.

As in our case there was an intersection of data on the scientific production of the departments with the present situation of the collection in itself, it is possible to identify and represent the core of knowledge of a particular thematic area, the peripheral subjects of knowledge, and the degree of interdisciplinarity.

With this visualization of the state of collections, information science professionals can make well-informed decisions about acquisitions and withdrawals. This should be done with a consideration of individual library needs as well as with a cooperative sense of the needs of the university library network as a whole.

References


1. Introduction

In this age of world wide communication, and the resulting thrust towards universality, the domain specific specialized vocabularies used in the visual resources and bibliographic databases pose limitations of access. Although they provide adequate access to the experts in the field, a layperson with little or no knowledge in the field (lay user) is constrained by the terminology. The meaning of the index terms used in the databases or the choice of search terms is not easy for such a user. It presupposes the ability of the users to demarcate and discriminate a concept/entity from other entities, and a lay user is often unclear about the connotations of the terms, and the subtle differences in their usage within the domain.

The challenge to any indexer of visual resources is the subjective nature of the interpretation of the image itself. Images very seldom contain any textual information besides title, creator or date. However, images also need to be accessed for content and context as well, and context and content description are very often left to "the eye of the beholder." Utilizing standardized terminology can help classify and define an image more accurately, thereby improving accessibility. However, the special needs of a lay user still remain to be addressed. The use of as an auxiliary resource may help alleviate this problem to some extent. This paper presents the findings of an exploratory study to assess the value of WordNet as a pre-search tool to aid in the understanding and the identification of concepts, including the terminology needed to search visual resources databases in the field of architecture. It also reveals the nature of the problems encountered by lay users and how specific aspects of WordNet were helpful.

WordNet is an electronic lexical database based on psycho-linguistic theories of human lexical memory, developed and maintained at Princeton University since 1985. Sets of synonymous terms (or synsets) constitute its basic organization. English nouns, verbs, adjectives are organized into synsets, each representing one underlying lexical concept. Several types of relationships between the synsets are recorded in Wordnet.

The users' study was conducted using the visual image database developed at the Architecture Library, Rensselaer Polytechnic Institute in Troy, NY. The database was developed using the Visual Resources Association's Core Metadata and the Getty's Art & Architecture Thesaurus for the descriptive terminology.
2. Overview of the method

The study was intended in an exploratory way to address the following questions:

- Does the use of WordNet facilitate searching image databases; aid in determining the “aboutness” of images; and assist in their description.
- Are there particular features of WordNet that are especially useful in this process?

This exploratory investigation utilized a multi-phased approach: (a) The observation and video taping of users while they performed designated tasks during their search for images in the RensSearch database. During these searches they used WordNet to help perform these designated tasks. Their computer sessions and vocal reactions were recorded using Camtasia \(^2\) video capturing software for subsequent review. (b) After each one of the online sessions, participants were given a written survey to complete which asked questions specific to their reactions to that particular exercise. In all, there were three sessions. (c) At the end in an interview, the users were asked to describe their reactions to the entire experience. Twenty undergraduate, graduate students, office staff and persons independent of Rensselaer participated in the study.

Session One: Finding Search Terms. Session one served to acquaint the user with WordNet and was designed to test its usefulness as a search tool. A list of eleven art and architecture-related terms were presented and from that list the user had to choose four to search RensSearch database. As the list of terms was non-index terms, the participants had to find alternative search terms in WordNet in order to locate images in the RensSearch database. For example, one of the listed words was mural and in order to find an image in the database they would need to search WordNet to find an alternative term such as fresco or wall painting to find an image.

Session Two: Aboutness. Session two was designed to test the usefulness of WordNet at a conceptual level. Participants were presented with six pages of images from which they had to choose three. Each of these pages contained five or more images of buildings or works of art that were similar in their shape, function or technique. They were asked to look at each set of images carefully and to choose a word that would effectively caption the “aboutness”, similarity or commonality of the subject images. Sometimes that commonality was functional, sometimes it was structural, and in other examples it was material. For example, one page contained a group of images that included a synagogue, temple, mosque, shrine and a church. The optimal term for describing these photos would be religious buildings or places of worship. The participants captioned the groups of images they chose and then searched RensSearch to see if they could find an image with a matching index term. If they didn’t find an image they then went back to WordNet and searched it to see if they could redefine, expand or narrow their original caption to find a term that more adequately described what that image set was “about”.

Session Three: Describing Specific Images. Session Three was designed to test the level of exactness or subject specificity of WordNet in helping a user find particular images in an art and architectural database. Participants were presented with fourteen images from which that had to choose four to search. The images included buildings, structures, drawings, paintings and spaces, both exterior and
interior. They were asked to find the specific images that they chose in the database and were told that if they didn’t find the image by using their own generated terms then to use the WordNet to help them find alternative terms.

Concluding Interview. The concluding interview was designed to elicit responses pertaining to the participants’ personal experience with the entire online searching process. They were asked to describe the usefulness of specific WordNet features, such as hypernyms, hyponyms, meronyms, synonyms, coordinate terms and the option to display the different senses of the term. Specifically, how these features helped them with terminology, clarification of ideas, narrowing down larger concepts and making-sense of the concepts and terms. If the participant did not find WordNet useful they were asked to describe in what way it didn’t help them.

3. Key issues and summary of findings

The following provides a summary of findings and a discussion of the issues that emerged from the study. It provides insight on the varied ways in which users may look for a work of art and the value of WordNet as a presearch tool.

3.1. Difficulties finding search terms

Participants were asked during session one if they had any difficulties finding search terms and whether or not Wordnet helped find search terms. 90% responded that it did provide alternative terms; 80% thought that it provided additional terms; 75% claimed that it helped them in making sense of the term and 70% said that it helped to refine the search term.

The analysis of session two revealed the mental strategies that users employed while attempting to identify the “aboutness” and a common caption for a group of images. These are not necessarily mutually exclusive strategies nor were they applied in any particular sequential order. Some of the predominant approaches included:

- Faceted approach: This involved using different aspects of the image such as; form (type of building, functions (usage), geographic location, time period, etc.
- Commonality strategy: Respondents looked for similarity among familiar images to find a caption term.

![Figure 1 - WORDNET as a Source of Search Terms](image)

3.2. Mental strategies used to determine group image captions

- Faceted approach: This involved using different aspects of the image such as; form (type of building, functions (usage), geographic location, time period, etc.
- Commonality strategy: Respondents looked for similarity among familiar images to find a caption term.
Specific to general approach: Respondents formulated generalized concepts by finding specific image attributes.

3.3. Difficulties in determining the group image captions

The nature of difficulties encountered by the respondents and the ways in which *WordNet* provided assistance is presented below:

- Lack of knowledge: 40% of the respondents had problems due to lack of knowledge of the subject matter.
- Choosing term to describe images: 80% of the respondents had problems with terminology and the following comments reveal how some were aided by *WordNet*:
  - "I didn't know synonyms, WordNet provided them."
  - "It gave me options of what to choose."
  - "Helped find connecting terms."
  - "Gave broader definition of term and related terms."
- Difficulties conceptualizing commonalities: 35% had difficulty finding a common theme amongst the images.

3.4 Describing Individual Images

The findings of session three revealed that 30% of the participants had problems conceptualizing the image. 55% had terminology problems such as, determining the appropriate keywords to use, finding specific terms, and narrowing down the terms, etc.

The *WordNet* assisted 70% participants in making sense of the terms or images, especially for those that lacked sufficient subject knowledge. With regards to choice of terminology, 80% found *WordNet* useful in providing synonymous, coordinate and hyponymous terms. As expressed by one respondent “visit the synonym city”.

![Figure 2 - Problem Comparison With Regard To Aboutness vs. Description](image)

Figure 2 presents a comparison of the two major problems, conceptual and terminology that participants encountered while performing the tasks in session two and three. Problems with terminology remained the more significant issue.

3.5. The Final Interview

The final interview was structured as a checklist of questions asking for the participants' views on the usefulness of specific *WordNet* features. Analysis of the responses indicated that it helped in several ways including; setting the term within a larger context (80%), clarifying ideas (65%), narrowing down concepts (80%),
identifying alternative terminology (80%) and making sense of terms and concepts (75%).

An interesting aspect of the findings indicates that participants felt that the most useful WordNet features were hyponyms, synonyms and coordinate terms. One of the negative aspects frequently cited was the confusing countless options it provided.

3.6. Observation of Search Moves Captured by Camtasia Software

The users' search moves and vocal comments were captured by Camtasia video software and were examined for patterns of search behavior. When participants were asked to determine the 'aboutness' of images they frequently used co-ordinate terms and synonyms. In contrast, for describing specific image content, they used the hyponyms and synonyms most frequently.

4. Conclusions

Evidence from analyzing the Camtasia videos and the results of this research indicate that it would be helpful to design a user interface capable of steering the users to a particular feature of WordNet based on the nature of the search. From these findings the conclusion may be drawn that WordNet is useful as a presearch tool in aiding the lay user in art and architectural searches of visual resource databases. Clearly, this topic merits further research.

References
Http://www.cogsci.princeton.edu/~wn/
Http://www.camtasia.com
Abstract: This paper discusses two inter-related themes: the retrieval potential of faceted thesauri and XML representations of fundamental facets. Initial findings are discussed from the ongoing 'FACET' project, in collaboration with the National Museum of Science and Industry. The work discussed seeks to take advantage of the structure afforded by faceted systems for multi-term queries and flexible matching, focusing in this paper on the Art and Architecture Thesaurus. A multi-term matching function yields ranked results with partial matches via semantic term expansion, based on a measure of distance over the semantic index space formed by thesaurus relationships. Our intention is to drive the system from general representations and a common query structure and interface. To this end, we are developing an XML representation based on work by the Classification Research Group on fundamental facets or categories. The XML representation maps categories to particular thesauri and hierarchies. The system interface, which is configured by the mapping, incorporates a thesaurus browser with navigation history together with a term search facility and drag and drop query builder.

1. Introduction

Our research investigates the extent to which a search system can play an active role in the retrieval process by automatic traversal of the semantic relationships in a thesaurus. Rather than relying on an exact match of terms, semantic term expansion makes possible flexible matching on 'close' terms. This offers various possibilities for retrieval tools beyond interactive browsing of the hierarchies. Items can be ranked by degree of similarity in a result or destination set, the system might automatically suggest terms to be considered for inclusion in a query, or having found one item of interest, a 'More like this' option can yield a set of similar but not necessarily identically indexed items.

Facet analysis is a key technique in knowledge organisation; concepts are decomposed into elemental classes, or facets, which form homogenous mutually exclusive groups. Faceted systems include MeSH, BLISS, PRECIS and the one discussed in this paper, the J. Paul Getty Trust's Art and Architecture Thesaurus (AAT). Recent times have seen interest in applying faceted approaches to the web and online retrieval generally (Broughton 2001; Chan et al., 2001; Pollitt, 1998). HIBROWSE (Pollitt, 1997) demonstrates the potential of faceted thesauri for information retrieval; hierarchies can be browsed and terms from different, "mutually constraining" facets interactively combined, with the number of postings shown dynamically as the query is constructed and refined.

The work discussed here seeks to take advantage of the structure afforded by faceted schemes for multi-term queries and flexible matching. Frequently queries contain multiple terms. Multi-term descriptors have potential for very specific item descriptions and high precision in retrieval. However, their full potential has yet to be exploited. The lack of flexible retrieval tools that can deal with ranked matches
of sets of terms is a disincentive to use of multi-term descriptors in indexing and search: "The major problem lies in developing a system whereby individual parts of subject headings containing multiple AAT terms are broken apart, individually exploded hierarchically, and then reintegrated to answer a query with relevance" (Petersen, 1994).

With the aim of increasing the universality of the work, our intention is to drive the system from general representations that can be mapped to particular thesauri and collections. To this end, we are developing an XML representation of important aspects of thesaurus structure and key system parameters. We report here on initial findings from this work, part of the ongoing ‘FACET’ project at Glamorgan (FACET, 2002), in collaboration with the National Museum of Science and Industry, which includes the National Railway Museum (NRM) whose collections database forms the dataset for the project. The system is implemented on a SQL-Server platform with a C++ retrieval engine and a Visual Basic interface incorporating a thesaurus browser with navigation history and ‘Favourites’ together with a term search facility and drag and drop query builder (Figure 1). While the thesaurus is stored as relational tables in the database, a parallel in-memory representation of the semantic network of relationships underlying the thesaurus permits real-time semantic term expansion, as outlined in Section 2. The main thesaurus in the project to date is the AAT (AAT, 2002, Soergel, 1995), although we are also incorporating smaller more specialised thesauri such as the NRM’s Railway Terminology Thesaurus. The AAT is a large thesaurus (over 120,000 terms). Descriptors are organised into 7 facets (and 33 hierarchies as subdivisions): Associated Concepts, Physical Attributes, Styles&Periods, Agents, Activities, Materials, Objects.

2. Semantic Term Expansion

Semantic term expansion is based on a measure of distance over the semantic index space formed by thesaurus relationships. Different cost factors are associated with traversal across the different relationship types, also taking into account depth in the hierarchy. Starting from a given term, the algorithm spreads over connecting relationships until a ‘semantic closeness’ threshold is reached (Tudhope & Taylor, 1997). Measures of distance between queries and item descriptors with multiple terms pose particular problems. The number of index terms may differ and terms may not match exactly. Here we outline work on a multi-term matching function that includes semantic term expansion (for details of the algorithm see Tudhope et al, 2002). Figure 2 shows the interface main window with list of queries and ranked results for the selected query. The top ranking result has 75% match since three out of four terms match and there is no penalty for extra item descriptor terms, whereas the 50% matches have just two matching terms. Item 1988-7335 has a higher match than 50% due to the additional partial matches on Carver chairs and Queen Anne Style. Similarly, item 1986-7794 fully matches on two terms but is ranked above the 50% items due to a partial match between upholstering and buttoning after semantic term expansion. For example, Edwardian, Regency and Victorian belong to a mini-hierarchy in the
**Styles&Periods** facet and *Queen Anne Style* is also a modern British style period. The bolded term *(armchairs)* in Figure 1 indicates a 'focus term' which must yield a match (after expansion). The choice of focus term can be dynamically changed with corresponding effect on results. Instead of one unified ranking, it would also be possible to produce rankings by facets separately. Queries and their result sets are stored persistently, using an XML representation.

![Figure 1: FACET interface - Browser and Query Builder with query used in Section 2](image)

![Figure 2: Queries with results (selected query: leather Edwardian upholstered armchairs)](image)
3. Fundamental Facets and Categories

The faceted approach to subject analysis began in 1933 with Ranganathan's Colon Classification (Personality, Matter, Energy, Space and Time). Ranganathan's analytico-synthetic approach was subsequently elaborated by the (British) Classification Research Group (CRG) who evolved an extended set of commonly occurring fundamental categories (Table 1) from the experience of developing a series of special classifications in the 1960s and 1970s and more recently BC2, the revised general BLISS Classification (Broughton 2001; McIlwaine and Broughton 2000; Vickery 1960). Broughton (2001) discusses the advantages of a faceted approach over enumerated classifications: avoids having to preassign all compound terms to the classification; avoids arbitrary assignment of complex subject headings to a single place in a hierarchical classification (or complicated polyhierarchies); subject headings can be synthesised from underlying facet elements. *Entity* is the main focus of an application domain – it could be types of mammal in a zoological taxonomy, subjects in library applications and in museum collections it will tend to refer to objects. It should be emphasised that facet analysis is more a technique than any one prescribed structure and that the CRG categories are not meant as absolutes but to be adapted for different applications. However the categories appear to overlap with several faceted schemes (eg Aitchison, Gilchrist & Bawden, 2000, p. 70) and capture useful higher level concepts. The current ISO standard also mentions common categories and work on the semantic network for the UMLS metathesaurus project (UMLS, 2002) has produced a set of categories suitable for various (medical) applications.

4. XML mapping of categories

Various projects have investigated XML thesaurus representations (eg Howarth 2000; Mitchell & Vizine-Goetz 2000; VOCML 2000). For our purposes, we wanted a level of representation more general than specific facets in particular thesauri or classifications. We wish to facilitate mapping from (eventually and as far as possible) a common query structure and interface to different thesauri and to different parts of the same thesaurus for different collections or different user profiles. One aim is to be able to shield a user from some of the complexity of a large thesaurus and dynamically map to particular hierarchies for a particular collection or information need. Another aim is to allow a user to express queries using a standard terminology that can be mapped to different thesauri.

An external XML file maps a category to a hierarchy or concept in a particular thesaurus. For figure 1, the XML mapping reproduced the AAT facets. Figure 3 shows the query discussed in Sections 2 and 3 but this time the browser interface is driven by a different XML mapping based on the CRG categories. Whereas before the AAT *Components* hierarchy was located under the *Objects* facet, the CRG categories give more prominence to *Parts* as a top level category and *Parts* has been mapped to *Components*. The point is that different mappings are possible for different applications, collections or user profiles. We employed a
subset of the CRG categories but any standard set of categories is possible. The XML representation offers the potential for a Query/Indexing tool to call up just part of a large thesaurus at some stage in the index/search process where browsing might be offered. Another possibility would be for the categories to drive 'Wizards' that encouraged users to construct standard faceted expressions of queries or index descriptors.

Figures 4 and 5 show an example of how the mapping of categories can affect retrieval functionality. Figure 4 shows the Find Term utility with no restriction on the scope of the search within the thesaurus (note the detection of compound terms, such as *Queen Anne Style*). Figure 5 shows the same search with the *Parts* category selected and resulting shorter list of terms from the mapping to the AAT *Components* hierarchy. This could be extended to restrict the scope of
automatic traversals in a retrieval system component.

5. Conclusions and Future Work

Thesauri have traditionally been intended to be read by human indexers and searchers, either on computer displays or in print form. The growth of digital collections has resulted in increased opportunities for automated processing of thesaurus representations. A matching function incorporating semantic term expansion can produce ranked results for queries with multiple thesaurus terms where individual terms are expanded and recombined. This may be useful for complex queries on large thesauri where manual browsing may be cumbersome. A semantic distance measure can also assist interactive query refinement. To reflect the move to online thesauri and assist automatic processing, there is a need for revised thesaurus standards which explicitly represent thesaurus structure, since we cannot reply on human interpretation of context. For example, it should be possible to distinguish automatically between Hierarchical Generic and Hierarchical Instance relationships and perhaps different kinds of associative relationship in electronic representations (Tudhope, Alani & Jones, 2001).

The initial experiments with higher level external XML representations of facet structure discussed in this paper are part of a longer term project to model the underlying semantic categories of concepts following a facet analysis approach. In future work, we intend to extend the representation to thesaurus relationships and query descriptor structure. A cataloguing or retrieval system could have different XML representations for different facet structures. The aim is not to build into the system implementation a dependency on the particular facets of the AAT (for example). Rather, we attempt a more universal approach, by seeking to identify general categories that can be mapped to the facets and hierarchies of the AAT or other thesauri.

Acknowledgements

We would like to acknowledge the support of the UK Engineering and Physical Sciences Research Council (Grant GR/M66233/01) and the support of HEFCW for the Internet Technologies Research Lab. We would like to thank Helen Ashby, Ann Borda, Sarah Norville, Charlotte Stone and other staff from the National Museum of Science and Industry for their assistance and the J. Paul Getty Trust for provision of the AAT.

References

http://www.getty.edu/research/tools/vocabulary/aat/
http://web.glam.ac.uk/schools/soc/research/hypermedia/facet_proj/index.php


Organization of the Information about Health Resources on the Internet

Abstract: The growing of the World Wide Web and its unstructured nature makes searches difficult, and moreover the information published by web sites is not always guaranteed. Usual searching engines are not enough to satisfy the needs of users and ease access to credible information. This paper focuses on the design of a medical information retrieval system that uses the MeSH controlled language and the fields structured to organize the information, mostly from Dublin Core metadata format. We use a set of criteria to guarantee the minimum level of information quality and credibility. The objective of such a system is to provide an efficient mechanism of information retrieval with greater searching capabilities than other subject gateways or search engines on account of its structure.

1. Introduction

An unquestionable fact is the massive and continuous growth of the information resources available on Internet, and especially on the World Wide Web. The facility to publish and the lack of control of the information are factors that have favoured the explosion of information on Internet which has become the greatest store of world-wide knowledge and, with no doubt, in the most used information sources, mainly being free. Nevertheless, the information that we can find can be of many kinds and created for multiple purposes, existing in different degrees of quality and reliability, due to the lack of screening and control.

These previous lines summarize the informative conditions of Internet: excess of information and questionable quality. This situation is generalized in any scope and discipline, but in the field of the health it acquires a great importance specially for two reasons; first of all, it is probably the field where most information is published; secondly, it has been demonstrated that the misinformation can be harmful for patients who use Internet as an alternative to find information on the treatment of their diseases (Kiley et al., 2002), and that although this information even comes from reliable and credible resources can not be correct for certain kinds of users (Impicciatore et al., 1997).

Under this situation, healthcare professionals and users face new challenges to locate useful, reliable and quality information. In order to try to solve the quality
aspect, different initiatives have arisen to develop ethical norms and to regulate the information quality, although it has been demonstrated that there are important gaps that must be treated in future (Risk, 2001). The main difficulties to evaluate the information quality are derived from the complex and multidimensional nature, while subjective, of the very concept of "information quality", and from the developed mechanisms and instruments to carry out the evaluation. Although different review sites have developed and used instruments to measure the information quality, it has not been possible to verify the construct validity and the interobserver reliability and its systems of rating are incomplete and of questionable utility (Jadad, 1998; Gagliardi, 2002).

In addition to the quality, the other important problem is how to search and locate relevant information. Although there is a great amount of directories and search engines, not only general but also specialized in health subjects, they lack of a clear and organized presentation of the information adapted to medicine and health world, which limits the utility of these tools in spite of the great number of resources that can contain. In order to solve this problem, several gateways of quality medical resources on Internet have arisen in the last years among them it is possible to emphasize in English language OMNI, Medical Matrix, Health on the Net, CliniWeb, Healthfinder, etc., and in French language CISMeF.

The indexing and organization of the resources in these subject gateways is based, generally, in the use of the controlled language MeSH (Medical Subject Headings) of the U.S. National Library of Medicine. Nevertheless, this is not enough to search for and to efficiently retrieve the information. Berland et al. show that the access to health information using search engines and simple search terms is not efficient (Berland et al., 2001), and with the exception of CISMeF, the possibilities of retrieval by different fields in the data bases of these gateways are very limited.

2. Objectives

The objective of this paper is to show an alternative proposal to the existing search systems and to describe INDISALUD (Index of Resources of the Health on Internet in Spanish), an Information Retrieval System (IRS) developed in the frame of a research project granted by the Vicerrectorado de Investigación of the Universidad de Zaragoza with the following purposes:

To structure, organize, describe, classify and index the health web resources in Spanish in a standardized and suitable way.

To expand and improve the retrieval options offered by the search services in the Web.

To provide enough information so that the professional and end-user can evaluate quickly his pertinence and some aspects related to quality.

3. Methods

To create and design the Information Retrieval System we have considered the following aspects:

a) Definition of the resources selection and inclusion criteria, in such a way that they act as a filter and they guarantee his potential relevance and credibility. In addition to the subject scope and resources language, criteria of inclusion and
exclusion have been defined related to the information reliability, adapted from the 
one's used in OMNI (BIOME, 2001), whose observance is mandatory to add the 
resources to the database.

b) To define the database structure: the information is organized and 
structured in 22 fields, 15 of them corresponding with the Dublin Core metadata. In 
Table I the fields are described; to assure the information consistency, some fields 
take their values from controlled lists.

c) To create a database with the resources candidates to be introduced in the 
INDISALUD system obtained by a search in several engines not only general but 
also specific, asking for the terms in Spanish language “salud” (health) or 
“medicina” (medicine). Once the URL addresses have been obtained, the following 
step is to identify the main address and to eliminate the repeated resources or those 
that belong to the same site. The resources thus obtained are those that have been 
verified with the inclusion criteria. Finally, this list closed when the 2,000 valid 
resources have been reached, the amount with which the service will be started 
initially.

d) The system input consists of the resource analysis and description 
according to the structure presented in Table I. The most relevant aspects 
correspond to the subject indexing and the abstracting, tasks developed traditionally 
by documentation professionals with the mission to represent the content of 
documents to be able to retrieve them later.

The subject indexing is made using as much descriptors of MeSH controlled 
language as terms of natural language, which facilitates the content representation 
in a consistent way and the exhaustive and accuracy retrieval of information. The 
fields in which the thematic subject of the resources is indexed has been organized 
hierarchically in several levels: medical specialty, descriptors, subheadings, 
identifiers and resource type.

These input tasks are developed by students of Librarianship and 
Documentation previously trained not only in the Web resources description and 
evaluation but also in techniques of indexing and abstracting and in the use of the 
MeSH controlled vocabulary. Later, the introduced resources are reviewed by 
documentation and medicine professionals, that are also in charge of the translation 
of the MeSH headings and subheadings into Spanish, avoiding therefore the 
possible variations in the translation and guaranteeing their consistency.

e) Retrieval mechanism: the search and location of resources can be made in 
two different ways:

- A browsing system through an alphabetical classification and a thematic 
classification of the resources, structured in five hierarchic levels: 
category of specialties, MeSH descriptors, subheadings, identifiers and 
resource type, which allows us to expand or to reduce the search by 
anyone of these levels.

- Search interface: its design allows us to enter the search terms directly, 
being we able to make the search in all the database fields or limiting it 
to certain fields, and to select the terms by consulting the index files. The 
search features include the use of Boolean operators, proximity 
operators, truncating, rank of dates, and explosion of terms in the fields 
of descriptors and resource types.
<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Resource name</td>
</tr>
<tr>
<td>URL</td>
<td>Uniform Resource Location (URL) Address</td>
</tr>
<tr>
<td>Author</td>
<td>Author/s or person/s responsible of the resource intellectual content. It can be a person, organization or service.</td>
</tr>
<tr>
<td>Organization</td>
<td>Author's institution or organization, including the department (if exists).</td>
</tr>
<tr>
<td>Editor</td>
<td>The responsible entity for making the resource available on the Net in its present format.</td>
</tr>
<tr>
<td>Country</td>
<td>Country where the resource is. Controlled list.</td>
</tr>
<tr>
<td>Province or State</td>
<td>Province or State where the resource is.</td>
</tr>
<tr>
<td>City</td>
<td>City where the resource is.</td>
</tr>
<tr>
<td>E-mail</td>
<td>Author e-mail of resource or publisher.</td>
</tr>
<tr>
<td>Abstract</td>
<td>Descriptive summary of the resource content.</td>
</tr>
<tr>
<td>Specialty</td>
<td>Medical specialty or specialties to which the resource belongs. Controlled list.</td>
</tr>
<tr>
<td>Descriptors</td>
<td>Subject indexing with the MeSH descriptors. The most specific descriptors will be used. In case it is required, MeSH subheadings will be included, added to descriptors. Controlled list.</td>
</tr>
<tr>
<td>Identifiers</td>
<td>Descriptors or identifiers of the resource not included in the MeSH. They will be used to describe the content in case there are not precise MeSH descriptors.</td>
</tr>
<tr>
<td>Resource Type</td>
<td>Type, nature or category of the resource. Controlled list.</td>
</tr>
<tr>
<td>Format</td>
<td>The physical or digital manifestation of the resource. Controlled list.</td>
</tr>
<tr>
<td>Source</td>
<td>Reference from where the resource is derived.</td>
</tr>
<tr>
<td>Creation date</td>
<td>Date when the resource in its present format has been created.</td>
</tr>
<tr>
<td>Update date</td>
<td>Last date of update of the resource.</td>
</tr>
<tr>
<td>Consultation date</td>
<td>Date in which the resource has been consulted.</td>
</tr>
<tr>
<td>Check date</td>
<td>Date in which the resource has been reviewed for the last time.</td>
</tr>
<tr>
<td>Audience</td>
<td>To whom the resource is directed, professionals or public in general.</td>
</tr>
<tr>
<td>Status</td>
<td>Active or inactive.</td>
</tr>
</tbody>
</table>

Table 1. INDISALUD database structure

f) Results display: the retrieved resources through the search interface show up sorted by update date or creation date, and it is possible to display them in two kinds of formats:

- Short format: it includes the title, URL, author, abstract and descriptor fields.
• Full format: it includes all the fields with the exception of consultation date, check date and status fields.

4. Results and Discussion

The result is an Information Retrieval System with Spanish web resources related to health aimed not only to the healthcare professionals but also to the public in general. Its main characteristics lie in the conceptual model of the information structure, the added value information from the abstract and the indexing using the most used controlled language in the medicine field, and the different search possibilities in every field with the consequent advantages of improving both recall and precision. All it conforms an efficient Information Retrieval System that differs more or less with the existing search engines and subject gateways; related to the first one the differences are derived specially as far as offering a structured right information and to the fact that the indication is done by people with a medicinespecific controlled vocabulary; related to the second one, because almost always they only offer a box to enter the search terms and they lack of options like search for in all the fields or to use the terms.

Another important utility of INDISALUD is that the user has access only to reliable and credible resources since they must fulfill inclusion criteria that act as a filter. The lack of a validated and reliable evaluation instrument, as well as low values in the indices of concordance between the judgments made by medical experts in the evaluation of the resources quality in the Web (Craigie et al, 2001), and the thought that the last judge in assessing the information usefulness and quality is the end user, has made us desist from carrying out an evaluation of the resources quality. Nevertheless, with the structure proposed we contemplate certain dimensions and characteristics related to the information quality that can be evaluated quickly by the user:

• Fast identification of reliable and credible resources by means of "author", "organization" and "editor" fields, that they provide information on the authorship and authority criteria of the origin of the resource, and the "source" field for those resources that have been published in printed medium and that have gone already through a review process.

• Selection of resources aimed at a determined audience through the "audience" field, in such a way that in the search those resources directed to professionals or interesting only for patients or users in general can be discriminated.

• Identifying the currently and update of the information through the "creation date" and "update date" fields.

• Selecting and evaluating the potential relevance or pertinence of the resources through the subject fields like "title", "abstract", "speciality", "descriptors" and "identifiers", as well as from "resource type".

INDISALUD will become accessible to the public in September 2002. From this date resources will be added to the system coming from both the periodic search in search engines and the sending by webmasters. The challenge is to
continue growing in the number of offered resources, but especially to maintain the database updated. Verifying the operativity of the resources periodically is simple and automatic; more effort implies to verify that a resource has been updated, since it will not only affect the update date but very specially the thematic fields like title, summary and descriptors that in case it is necessary will have to be modified to maintain their informative value and to reflect the resource accurately.

5. Conclusions
The great amount of existing information on Internet, the different types of formats, the lack of structure and the variable degree of quality are some of the factors that cause the valid and useful information search in this environment to be difficult. As happened fifty years ago with the phenomenon of the scientific information explosion that gave rise to the appearance of the online systems with their referential or full-text databases, it becomes necessary at the present time to develop information retrieval systems, called thematic gateways or otherwise, that allow the search and efficient retrieval of relevant information and with certain quality and credibility assurance on Internet.

The efficient retrieval depends not only on the capacities of the search system and the searcher ability, but also on how the information is organized and structured and on the precision whereupon the resource content is represented during the description, indexing and abstracting process.

The structured information model proposed in this paper does not represent any new development in relation to the traditional bibliographical databases. It simply tries to know how to take advantage of the accumulated experience and knowledge throughout the years and to implement those characteristics of the traditional systems that still at the present time represents an advantage on the current systems, or in form of structure, metadata or retrieval languages.

Notes
1 A search made in Google (http://www.google.com) the 8th of March 2002 with the simple term "health" recovered 68,500,000 resources.

References


Gagliardi, A.G.; Jadad, A.R.; (2002). Examination of instruments used to rate quality of health information on the internet: chronicle of a voyage with an unclear destination.


Eduardo Péis and Antonio Ruiz  
Department of Library and Information Science, University of Granada, Granada, Spain  

Francisco José Muñoz Fernández  
European Documentation Center, University of Granada, Granada, Spain  

Francisco de Alba Quiñones  
Documentation Center, Tourism and Sport Office, Andalucia Government, Spain  

Practical Method to Code Archive Finding Aids in Internet  

Abstract: Actually information services cannot work ably without web functionalities. Archives are paradigmatic examples of information systems with low use of technological possibilities. A small but hopeful step forward is the electronic encoding of descriptive information. One tool made suitable for it is a DTD - SGML for instruments of archive description named EAD (Encoded Archival Description). Having a base in an international norm like SGML, this tool, makes the information interchange possible. In this work Microsoft Active Server Pages encounter a concrete method for the automatic labeling of instruments of file description utilizing the displaced to one side programming technology of the server. Although the method develops in aggregate utilizing Microsoft technology, it can be utilized in another platforms, since the result generated is a document labeled in ASCII, to plain text, and therefore easily transferable to other platforms.  

1. Introduction  

If archives were adjusted to the new technological possibilities, they might be exemplary in making efficient use of resources and increasing the degree of use of the information. One key step on this road to adaptation is the design of structured finding aids that would facilitate web diffusion of archival information, as well as the re-utilization of this material in the form of varied informational objects.  

To start with the process of adaptation, these functionalities can be developed by creating a "digital" version of the finding aids, encoding the structured descriptive information with a "mark-up language" that conforms to international standards.  

Precisely an international standard like SGML (Standard Generalized Markup Language) is the best option for conceptually codifying structural units of documents, in the broadest conception of the word, for many reasons. First of all, it is a generalized public standard, independent of manufacturers and distributors, which ensures its permanence. Secondly, it does not impose a fixed set of components: the structures can be defined by the user, under the concept of
“document type.” Therefore, it is remarkably flexible and consistent, adapting to whatever alphabetical system may have been used to write the text. And finally, it breaks down the structure of presentation, as it is a descriptive codification rather than a procedural one. For these reasons, more and more applications are using SGML as a basic tool (Cover).

It is easy to understand why the first attempt to apply a standardized code to instruments of archival description --initiated by the University of California at Berkeley and directed by Daniel Pitti-- selected SGML as the ideal technique for carrying out such a codification.

This project came to design a DTD (Document Type Definition) to describe a class of documents which consists of an optional title page, the description of a unit of archival material, and appendixes, which are also optional. The title page could include varied elements such as the identification of the repository or the type of descriptive tool. The descriptive component, in agreement with the DTD, would offer a brief description of the unit (using markable elements analogous to those employed in a MARC catalographic record), a broader narrative description of the unit and any of its separate parts (including markable elements such as title, dates, scope and contents) and a formatted list of the parts that contain that unit.

Logically at the present time the EAD DTD is usually developed in XML (eXtensible Markup Language). XML is a subset of SGML, but while SGML is mostly used for technical documentation and much less for other kinds of data, with XML it is exactly the opposite, being it more usable for distributing materials on the Web (Goldfarb and Prescod, 1998).

Therefore, as SGML, XML provides the rules for defining a markup language based on tags. It has been developed to keep up the proliferation of proprietary formats in use for electronic document processing and representation. It is a “descriptive” system that gives a declarative and machine-independent description of the document structure using codes that simply offer names to categorize and identify the parts of a document. This means that XML is a protocol devised to articulate structures of contents of documents instead of the appearance of documents.

The current EAD model (version 1.0) offers the option of using XML. It is possible to activate/de-activate the variant sections SGML/XML option using the SGML feature called “marked sections.”

Despite the apparent advantages of applying this model, many archival systems can run into formidable difficulties in implementing the XML descriptive system. Our proposal consists of the development of a customizable template which can computerise the process of input of the descriptive information, convert and adapt this information to the EAD model and later handle the informational XML object with complementary technologies.

2. Methodology

To develop this method we need a base in an existent archive, the archives of "Patria" newspaper, applying to the ISAD (G) standards. In this way, we can count on an accessible and close form of a archive description instrument, strengthened and made use of a real file, and we can apply the encoding methodology according to EAD directly, without the previous process of the construction of the description instrument.
The description comes true on three levels: funds, section and document. In regards to this communication, we focus on becoming centered in the document level, given the limitations in extension impede a complete analysis. In any case the methodology is the same on the distinct levels; only ISAD's fields vary in that they get into stake encoding.

Example: we present a document's concrete description, in a photo, following the structure ISAD (G):

1. MENTION AREA OF IDENTITY
   1.1. Referential code: ACT 1/C3
   1.2. Title: La casa de Ángel Ganivet
   1.3. Dates: 11 enero 1979 (fecha de publicación)
   1.4. Description level: Unidad documental
   1.5. Characteristic day girls: 13 x 18,5 cm, positivo, papel, B/N

2. CONTEXT AREA
   2.1. producers/autor name: Ferrer

3. CONTENTS And STRUCTURE area
   3.1. Catch up with and contents
   Photo of press. Casa-Molino de Ángel Ganivet, on the Cuesta de los Molinos (Granada)
   Photo caption: "Unico de los rincones granadinos de mayor sabor romantico es este de la Cuesta de los Molinos y se centra de manera especial en el rincÁn que crea la casa de Ángel Ganivet. Ahora se ha levantado la voz de alarma porque parecer la Casa del Molino de la Zagra, en la que esta la lápida con el busto del escritor y diplomÁtico granadino, esta en situaciÁn ruinosa y cualquier día se nos puede venir abajo a todos. Porque la casita deberÁa conservarse como reliquia que nos aproxima a la formidable figura del autor de "Granada la bella"

4. DOCUMENTATION ASSOCIATED AREA
   4.1. Publication note
   5. FotografÁa publicada en el Diario PATRIA el 11 de enero de 1979, p.20 (contraportada)

6. NOTES AREA
   6.1. Note: Se ha tomado como tltulo el que acompafia a la fotografa publicada.

KEYWORDS / INDEXING TERMS
   URBANISMO, INDUSTRIAS, CASAS-MOLINO, GANIVET, ÁNGEL, CUESTA DE LOS MOLINOS, GRANADA.

Anyone can notice that there is information that is not described, but it was done in this way because this information appears in finding aids on superior levels. Granted that this description instrument is already in use, we have a record data base with descriptions of distinct documents. We for applying automatic encoding to this data base. Thus, our objective will be to have a data base, and to encode it with EAD in time of to take it to Internet and as about this business in automatic form.

There can be several forms, but the method begins with a same starting point: The equivalency among the ISAD fields and the EAD labels. Obviously, it would be possible to construct independent labels from XML, but then we would not get this encoding's other objective: normalization. In this way, we would begin by constructing a document

<?xml version="1.0" encoding="ISO-8859-1"?>
To accomplish the automatic codification process, we used the Active Server Pages technology of Microsoft, a programming language of the server's side geared for Internet Information Server and Personal Web Server. The reason for selecting this option is its implementation facility in Windows systems, as well as the accessibility of most of the archives without having to resort to more complicated systems.

Thus, departing from the base of relational data base where the records are stored with the descriptions ISAD (this data base can be Access, SQL Server or any other accessible by means of ODBC), encoding would be effected. Then several alternatives can be presented. A first would be to connect the data base with a Web fill-out form, to consult the data base with sql entences. Thus, in our project we created a fill-out form to consult the data base, and the results will format themselves using xml, that is, the file will be in asp language but the contents would be completely xml; with the distinct EAD labels to encode contents. Asp language permits this possibility, though right now an asp file is no more than a file with format labels and code labels that must process the server. This way we can include labels as xml without problem: the processing instruction of document xml are given and the patch of IE5s shows it on the screen correctly. In this way, the automatic encoding of the description instrument for the distinct records is effected on the one hand, and in addition, the same ones are shown in an easily exportable format, in plain text, totally compatible text with any other system that follows EAD DTD without errors. The xml and css style sheets combination and the use of name spaces allow us the appropriate data presentation in the usual format of the current
web sites. This method avoids the IE5 default presentation in form of pull-down nodes. Also possible is the use of introduction data technology to include the xml code with the search results among the html code. This second option improves on the previous one in that the xml code is clearly defined, facilitating the exchange between systems. The xml code is not mixed with labels of different name spaces and the identification of the EAD content is easier.

A second alternative is the use of a complete database in one xml file. The first step is coding the complete database in xml files following the label EAD list. This is an automatic process that we did using the same previous method, but now without a form to select records because we code the whole database. In this case, the selection of records is carried out using xsl, with which we can control the localization of the documents EAD nodes, as well as place conditions for data output. To have the database with EAD labels and in plain text format allows use of the DOM technology to edit and change a EAD document. With this method, we can control from one page web different types of data, applying different xsl templates to our EAD label database. Here more programming is needed to show the EAD records in the web, but we have more compatibility between different systems.

3. Conclusions

It is absolutely necessary a normalize in the handling of archive information to face to the new challenges that are presented by diffusion of information in Internet.

The EAD DTD is a xml document type definition that results very appropriate to encode archive finding aids according to the ISAD standard.

It is possible to devise a customizable input template which allows automatic EAD encoding of archivist finding aids.

The use of server programming technologies like active server pages allow the code of databases made according to the ISAD standard. These technologies allow data output in a EAD DTD labeled format, guaranteeing a higher compatibility between systems.

The trend to EAD document database models is a way of assuring the compatibility of the current information in the new systems, and can be a more effective form of freeing data of applications, avoiding software obsolescence derived problems.

We should consider that the application of the EAD is more transcendent than the application of the ISAD (G) standards, since the electronic document is today a reality in the archive.

References
Investigation project for the recovery of the "Patria" newspaper photographic archive carried out by Dña. Antonia Salvador directed by Dr. Antonio Ángel Ruiz.

Berkeley University EAD web page http://sunsite.berkeley.edu (visited April 16, 2002)
Abstract: The purpose of this paper is to explore the feasibility of using a business process model as a framework for the integrated organization of electronic information in the context of business enterprises in general, and more specifically in the SMME sector in South Africa. An SMME is defined as a separate and distinct business entity, managed by one owner or more, and having less than 100 employees. Information organization in the business environment is described within the framework of three contexts, namely (1) document creation, (2) collections of documents, and (3) information retrieval systems consisting of document surrogates. Internal and external information resources typically found in business enterprises are categorized according to their relationships with business processes. It is concluded that it is theoretically feasible to organize resources in SMMEs according to a business process model.

1. Introduction

The purpose of this paper is to explore the feasibility of using a business process model as a framework for the integrated organization of electronic information in the context of business enterprises in general, and more specifically in the SMME sector in South Africa. At the same time it is hoped that this approach of relating information organization to business processes will demonstrate the potential value of information organization in the business environment. It should be recognized from the outset that the owners and managers of enterprises are primarily interested in the survival and growth of their businesses. Therefore, if the information professional wants to prove the value of an activity such as information organization, that will require resources (human resources, finance, equipment) to be made available, it will be necessary to show that this activity supports business processes and strategic goals, and thereby contributes to the competitive advantage of the enterprise, even if only indirectly.

This paper reports on some of the work done during the first year of a larger project scheduled for at least three years, and covering all aspects of the organization of information in SMMEs in South Africa. The focus in the research project is on SMMEs because we think that the need for advice and assistance in organizing business information is the greatest in this sector. Many large corporations recognize the importance of information/knowledge as a resource, make provision for its efficient management by appointing professionally trained and dedicated people as information/knowledge managers, and use sophisticated, expensive corporate information systems. Medium, and especially small and micro enterprises, however, usually cannot afford such luxuries. In many cases there also seems to be a lack of awareness at management level of the importance of
information as a resource, and a lack of knowledge about the technology and procedures that can be used to organize it. This situation leads to the development of filing systems and databases according to the preferences of individual workers. Where formal information systems are used they are often developed in a piecemeal fashion for the purposes of individual business processes or units such as finance, customer relations management or human resources management. These unintegrated, decentralised and individualised systems (or lack of systems), are obviously not conducive to the full utilisation of information resources for the benefit of the company.

A further reason for focussing on SMMEs is that this sector plays an increasingly important role in the South African economy in general and in the developing communities in particular. In 1997 96.5% of all business enterprises in South Africa were located in the SMME sector, and these businesses contributed 32.7% to the GDP (Martins & Tustin, 1999, 47). This has now increased to 42% (SA: an overview, 2002). It is also generally acknowledged that SMMEs make an important contribution to job creation, the redistribution of wealth and improvement of the competitiveness of local industries in the global economy (e.g. Martins & Tustin, 1999, p. 1; Rogerson, 1997, pp. 5-7). It should also be noted that in the developing communities nearly all businesses fall in the SMME sector. If information can indeed be organized in a way that supports business processes it would indirectly contribute to the survival and success of SMMEs and to development in South Africa.

2. Research questions

To investigate the appropriateness of a business process model for information organization in SMMEs the following questions will be adressed in this paper:

- Which businesses should be regarded as SMMEs for the purposes of the investigation?
- What does information/knowledge organization in the business environment entail?
- What are the typical business processes described in the literature?
- Can the information/knowledge resources that businesses might want to organize be related to business processes?

Many more questions are being addressed in the larger research project, but these fall outside the scope of the present paper.

3. What are SMMEs?

A number of criteria to be used in defining SMMEs are suggested in the literature: number of employees, annual turnover, credit requested from banks, total value of assets and annual balance-sheet total. The most generally applied criteria seems to be the number of employees and annual turnover. The SMME sector in South Africa is defined by the National Small Business Act (Act 102 of 1996) in qualitative and quantitative terms. The qualitative criterium states that to qualify as an SMME the business must be "a separate and distinct business entity ... managed
by one owner or more ...". Therefore the business should not be a branch or part of a larger organization, but a business in its own right with elements such as customers, suppliers and products. On the basis of quantitative criteria (employees, turnover and total gross-asset value) the Act distinguishes four categories of enterprises, namely micro, very small, small and medium. Rogerson (1997, p. 2) and Martins and Tustin (1999, pp.26-27) add survivalist enterprises as a fifth category, hawking and subsistence farming being examples. It can be assumed that survivalist enterprises generate and acquire very few if any information resources that would need organizing.

For the purposes of this project enterprises will be categorised according to the number of employees as micro (including the category of very small) (1-9 employees), small (10-49 employees) and medium (50-99 employees). Taking one or more of the financial criteria mentioned above into account would complicate the categorisation of enterprises because a business might fall into one category according to the number of employees and into another according to annual turnover, for example. It can also be argued that the financial criteria are not very relevant with regard to a company's needs for information organization, although finances will of course determine the technology (hardware and software) that the company can afford. On the other hand, the number of employees can play a significant role in the amount and diversity of information resources that have to be organized, and in the need for centralised and integrated information systems.

4. Information/knowledge organization in the business environment

The Library of Congress Subject Headings defines information organization as "identifying, describing, and providing access to information-bearing entities in all kinds of environments, such as archives, libraries, museums, offices, and on the Internet, through the gathering of the entities into organized collections, and/or through the creation of retrieval tools, such as bibliographies, catalogs, indexes, finding aids, registers, search engines, etc." In terms of this definition the present paper deals with information organization in the office environment, more specifically offices in SMMEs. According to Taylor (1999, pp. 13-14) information organization in offices takes place in manual and electronic records management systems, and involves the organization of units such as directories (folders), data files, computer programs and fields in records. These units have to be organized at enterprise level as well as at personal level by individual staff members. At the enterprise level information has to be organized in systems such as folders on network servers, intranets and corporate portals, document management systems, electronic mail systems, transaction processing systems, management information systems, executive information systems, decision support systems and groupware systems. Many of these systems are extremely expensive and will usually be found only in large corporations, not in SMMEs.

Based on the LCSH definition and authoritative handbooks on information organization, such as Rowley and Farrow (2000) and Taylor (1999) one can say that information organization takes place in three contexts, namely (1) in the context of the creation of documents, source databases and other information-bearing entities, (2) in the context of forming collections of documents, and (3) in the context of constructing information retrieval systems consisting of document surrogates (bibliographic records).
At the time of the creation of information-bearing entities the following information organization activities can take place in a business enterprise: the provision and encoding of embedded metadata in documents published on the company intranet; the structuring of information elements on a corporate website or portal; the organization of data elements in tables and fields in a relational database, e.g. a customer or product database; the construction of a knowledge repository for capturing the knowledge assets of the company; the construction of an index to a document, e.g. a project report, annual report or staff manual.

Organizing collections such as corporate or personal libraries and archives of printed or digital information entities entails the process of classification (categorisation) for the purpose of physical storage and retrieval. Classification can be applied to the shelving or filing of printed documents according to a published or home-grown classification scheme, and to the storage of digital objects on disk in some kind of classified folder structure.

The creation and organizing of document surrogates in bibliographic retrieval systems include the processes of description, indexing, abstracting and classification of information entities, resulting in systems such as catalogues, search engines, directories and subject gateways. Various tools (system aids), such as retrieval system software, codes for document description, metadata formats, indexing languages and classification schemes are used in creating these systems. The construction of bibliographic retrieval systems is probably not a high priority in most SMMEs, especially those in the micro and small categories, although they might be a necessity in information-intensive enterprises such as law practices.

5. Business processes

Alter (1996, 60) provides a useful model of generic business processes, grouped into three categories (terminology used by other authors are added in brackets where appropriate):

A. Processes requiring coordinated work from many functional areas

These are processes that cross functional areas. Specific processes mentioned by Alter are the creation of a new product, creating a coordinated plan for an entire business and fulfilling customers’ orders.

B. Processes typically related to a specific functional area

- Production (Operations): purchasing materials, assembling or fabricating the product, delivering the product, servicing the product and supporting the customer.

- Sales and marketing: identifying potential customers, deciding what method customers really want, identifying market opportunities, making customers aware of the product, persuading customers to buy the product, performing the sales transaction.
• \textit{Engineering (Research and Development)}: performing research about new methods, determining how to produce products, determining how to improve production processes.

• \textit{Accounting and finance}: performing financial transactions, creating financial statements, paying taxes, investing cash, financing operations.

• \textit{Human resources}: determining hiring requirements, hiring people, introducing employees to the company operations, paying employees, administering employee benefits, administering disciplinary actions and terminations.

C. Subprocesses and activities occurring in all functional areas

These pervasive processes include communicating with other people, analyzing data, motivating employees, planning work to be done, keeping track of work being done and providing feedback to employees.

6. Information/knowledge resources and business processes

The busy manager of a small enterprise struggling to survive might not be easily convinced that it is necessary to spend a lot of time and money on the organization of the company's information assets. However, if it can be demonstrated clearly that organized information is critical to the operations of the enterprise and supports the business processes he/she might be more willing to invest in systems aimed at the capturing and organized storage of the information. It is therefore necessary to determine whether information resources can be directly related to business processes. In this discussion the information resources will be grouped in the two broad categories of internal and external information, a distinction used by several authors (e.g. Choo, 1998, p. 139; Pollard, 1999, p. 89).

From the point of view of information organization an important question is whether external information resources should be organized as a separate collection or integrated with internal information. If the organization of both broad categories could be based on business processes, it might be feasible to create a single integrated system for all resources. In the following brief overview of internal and external resources an attempt is made to relate these resources to Alter's model of business processes.

6.1. Internal information resources:

This expression refers to all information relating to the internal environment of the company. Most of these resources are produced by company employees, and emanates from the business processes outlined above. Some resources originating outside the company, e.g. bank statements, accounts and receipts from suppliers, orders from customers, tax assessments, auditor's reports, etc. should perhaps also be regarded as internal information because of their direct relationship with the internal environment.

Many types of resources can be identified by simply analysing the list of business processes compiled by Alter and by consulting other handbooks on information systems. Cross-functional processes produce information resources such as product development reports, business plans and competitive intelligence reports. The production/operation process involves documentation on the
purchasing of materials and equipment (information sheets from suppliers, orders, bills, etc.), records of quality control, delivery notes, guarantee cards received from customers, servicing records and customer records. The sales and marketing function produces market research reports, product brochures and information sheets, advertisements, press releases, orders from customers, records of sales transactions. Internal information resources related to the engineering (or research and development) process include project planning documentation, laboratory notes and project reports. Accounting and finance give rise to budgets, regular financial reports on income and expenditure, documentation relating to taxes, records of investments and assets, etc. Information resources produced in human resources management include job descriptions, advertisements for vacancies, documents relating to employee benefits, employment contracts, training manuals and other training materials, employee records, including records of payment, leave and disciplinary hearings. In the course of the subprocesses and activities occurring in all functional areas information resources such as letters, memorandums, email messages, results of data analyses, schedules of tasks and project documentation are produced.

6.2. External information resources:

This category of resources include all information entities originating outside the enterprise and containing information about the external environment. There is general agreement in the literature on competitive intelligence (business intelligence, environmental scanning) (e.g. Choo, 1998; Pollard, 1999, Vine, 2000) that in gathering information for intelligence the focus should be on customers, suppliers, competitors, competing products and services, government policies and regulating instruments, and political, economic, social, demographic and technological trends.

Many information items gathered for the purposes of competitive intelligence support the cross-functional business processes, especially strategic planning and decision-making. In the creation of a business plan, which can include processes such as the setting of strategic goals, determining niche market segments and deciding about mergers with or acquisitions of competitors, the top management of a company have to rely heavily on external information resources. These include resources such as general news and economic journals, cuttings from national and local newspapers, trade journals, annual reports, planning documents and environmental filings of competitors, reports on court cases against competitors, publicity materials of competitors collected at conferences and trade exhibitions and land registry documents. In the production/operations process there seems to be heavy reliance on internal information and little need for external information, except perhaps descriptions of competitors’ processes as a source for benchmarking and improving the company's own processes, and documentation from suppliers of materials and technology. External sources needed for sales and marketing include market reports by market research companies and academic research units, geographic information (e.g. maps), directories for information about potential customers and evaluations of CRM software. In research and development the creation of a new product or the improvement of an existing product or manufacturing process involves extensive use of external information such as research reports, conference papers and patent literature. For accounting and finance external information is needed about tax law changes, sources of external finance (e.g. a small business development corporation or government funding), investment
opportunities, financial software, etc. Human resources management is dependent on a variety of external information sources such as labour legislation, trade union documentation, sources about industry trends in salaries and employee benefits, guidelines for performance measurement and video and audio tapes for training.

7. Discussion of implications for information organization

The categorisation of typical business information resources according to a business process model above suggests that it is theoretically feasible to organize resources in small, medium and micro enterprises (and larger enterprises as well) according to such a framework. These processes can for instance be used as the basis of a classified system of folders on a network server or on the hard disk of a personal computer, categories in a directory of resources on a company intranet, the structure of a corporate portal, etc. They can also provide the terminology for an in-house thesaurus for indexing to be used in a metadata system.

It is not suggested that the model of business processes used for analysis here, or any other theoretical model, be applied rigidly in the development of a system for classification or categorisation. The current trend is to move away from a strictly functional division of business activities and apply other approaches such as organizing around customer-oriented processes, geographic regions or product groups (Alter, 1996, 59). In a specific enterprise one should therefore analyse the actual structuring of business processes and use that as a model for organizing the information resources of the company. Such an approach that aligns the systems for information organization with the company's business processes should ensure the best chance of convincing the managers that information organization can support these processes, and thereby contribute to the strategic goals and competitiveness of the company.

Acknowledgements

The assistance of MA students Lize Denner and Nomcebo Magagula is gratefully acknowledged.

References


Ricardo Eito Brun
Carlos III University, Madrid, Spain

Software Development and Reuse as a Knowledge Management Practice

Abstract: Software development is usually referred to as a knowledge intensive practice. In fact, companies involved in software development are said to be the most clear example of the companies whose revenue and value is based in the creation, representation and modeling of knowledge.

The software development process starts with the capture of some requirements provided by the final users or by the project stakeholders. Taking these requirements as a basis, the development teams must cross the chasm from the text-based requirements to their representation in a specific programming language.

To move from the textual representation of the software requirements to the final code, different steps and artifacts must be created. These artifacts also represent the knowledge embedded in the functional specifications, and are an intermediate step to reach the final representation: the programming code.

The author analyses the impact of knowledge management in software development processes, and describes the possibility of using a shared pool of code to make possible the sharing of the knowledge gained.

A prototype tool to create and maintain the repository of code and link this code to functional specifications and any other artifacts (documents, models, reports, etc.) is described. This tool will link together the knowledge created at the different levels (business analysis, analysis, design, coding, etc.) in the development process to make the reuse of code easier.

1. Knowledge Management

Today, knowledge management (KM) has become a key issue for any organization regardless its size and business activity sector. Companies recognize KM as an important practice to achieve competitive advantages, as KM programs can give them the possibilities of:

a) increasing innovation; b) cutting the time needed to design and market new products and services; c) ensuring the effectiveness of your employees; d) reducing the costs of the training and external consultancy and e) incrementing the motivation and participation of the employees

This has led to an increasing number of projects focused on the development of KM strategies. But the first challenge companies must face when establishing a KM program is to clearly define the scope of the project and give answers to a sometimes difficult question: what do we mean when we are talking about knowledge management?

KM means putting in action different practices with a single purpose: make the creation and transfer of ideas and knowledge easier. More concretely, KM means:

a. Gathering best practices and the most efficient work procedures, and deploy them throughout the whole organization.
b. Sharing the ideas and the tacit knowledge available in the brains of the employees, clients and partners.
c. Making all this knowledge and expertise explicit.
d. Creating knowledge bases or repositories where the explicit knowledge is linked to work processes and daily tasks.
e. Ensuring that all the employees can access the knowledge they need to complete their tasks, work and assignments.
f. And finally, encouraging employees to use the knowledge, and measuring this reuse

2. Knowledge management and software development

Software development is usually referred as a knowledge intensive industry. In fact, when developing software, analysts and programmers must identify the knowledge about business practices provided by the stakeholders and future users of the systems, make this knowledge explicit by means of functional specifications or any other kind of documents, and transfer all this information to other people in the production chain that will take them as a basis to create additional artifacts until the final coding and deployment of the software.

During this process, knowledge must be made explicit and shared between people. Different artifacts are used to represent it: textual information, graphics and diagrams, formal languages, programming languages, etc. These artifacts can be seen as different levels of knowledge representation.

Most companies have adopted software development methodologies to provide workers with guidance and rules they must follow to make the transition from one level to another.

These methodologies try to standardize the best possible way of creating software and moving from the initial specifications to the final code.

However, the success of these methodologies is not clear. Some people are skeptical about them, and the software development process is still considered to be something difficult to manage: an anarchic process that depends, in a great extent, on the personal commitment and knowledge of the individual programmers.

People involved in software development consider these methodologies as an additional and unnecessary workload that do not supply any real improvement in the development of the applications.

These considerations are due to a lack of knowledge of the advantages that we can achieve in the near future once we have spent a little time representing knowledge. In fact, the importance of the methodologies goes beyond the document types they recommend to write down. What a methodology really provides is a systematic way to represent and organize both functional and technical knowledge.

In general, although methodologies are sometimes used to represent knowledge, little effort is currently being done to improve the sharing of this knowledge. So, people think of methodologies as overwhelming papers instead of a knowledge sharing tool.

As a result, in order to develop better software and help manage current and future development projects, some techniques and practices developed in the KM area should be borrowed and merged with the knowledge representation techniques proposed by the traditional software development methodologies.
3. The role of documentation in KM

The success of any approach aimed to improve the management of knowledge in a software development company will depend on our ability to:

a) Capture and make explicit the working knowledge created by the community of developers and organize it in an efficient way.

b) Provide context to this knowledge by means of different artifacts that cross the gap between the requested functionality, the code, and the intermediate documents created with modeling languages such as the UML (Unified Modeling Language).

c) Reward the reuse of knowledge.

Software documentation plays a key role in the proposed KM strategy. Although sometimes deprecated and considered an inconvenience, one of the main steps in any KM strategy consist of writing down the knowledge available in our brains.

This process, referred to as externalization in Ikujiro Nonaka’s classic book *The Knowledge-Creating Company* points out the need to write down what we know and move it to an organized and shared repository of documents.

This makes software documentation important, as it must be understood under a new perspective: documents should be seen as the tool we use to transfer knowledge and make this knowledge available for further reference.

4. Problems due to the lack of documentation

But regardless their impact on a KM strategy, the lack of documentation is the origin of several problems in a software development organization:

*Impression of low quality.*

Systems and system development processes cannot be audited if no design documents are available. External evaluators and inspectors consider the lack of documentation as a symptom of insufficient organization and inefficient work processes. If the system documentation is not up-to-date, the results of the audits and the subsequent evaluation will be negative to the system and the whole organization.

*Higher development and maintenance cost.*

Finding defects in the early phases of software development (that is to say, in the specifications of the requirements and the design instead of in the programming code), significantly decreases the costs of defect removal.


Without documentation, the only reliable and objective information is the source code itself. In such a situation, programmers must spend a great amount of time trying to understand the system's functionality by exploring its source code.

Experience shows that the time spent by maintainers studying source code is 3.5 times greater than the time they should spend studying documentation (Source SEI CMU).

So, documentation is critical for software engineers and technical managers responsible for the evolution of the software.

4.2. System integration.

Software applications are not isolated pieces; they need to talk to each other to better support critical processes. The evolution of IT makes unpredictable the
future requirements of any company, as well as the needs to interface with coming technologies and applications.

Documentation ensures the possibility of extending the system with interfaces to other applications provided by third parties. The lack of documentation will force the development teams to face costly re-engineering projects when an integration with other system is requested.

From a business perspective, the difficulty found when designing interfaces to other systems by a lack of accurate documentation also have a negative impact in the time-to-market and reliability of future IT projects. In fact, even the ultimate application just developed will become legacy in the near future.

4.3. Corporate amnesia.

Employees' turnover is one of the most important problems for software development projects, and although it is true that nobody is irreplaceable, there are costs associated with employee turnover. Errors made by novice programmers who do not have a detailed description of the software application reduce the efficiency of the whole system, and the quality of the code degrades as more and more people updates it without the guidance of documents.

Problems in team communication. In any software development project, communication is a key factor to obtain success. Documentation helps companies improve communication between the different teams and developers involved on the project. This factor is even more important in the case of companies and projects where external consultants or programmers are hired in a temporary basis.

5. Problems with the traditional approaches to software documentation and proposed solutions

There are several reasons that detractors of software documentation put forward:

a) Documentation does not make the code more maintainable.

b) Documents become obsolete as soon as they are created.

The reason of these problems are that usually documentation does not contain clear links between the functionality requested by the Client and the system code and technical descriptions. As a result, people need to invest a lot of time reading a huge amount of "paper", and even after reading this overwhelming documents, they cannot gain a clear understanding on how a specific functionality has been implemented, as documents are not related to each other or to the real implementation.

To solve these difficulties, we need to offer a documentation set that links functional specifications and list of features to the documents explaining their technical implementation and to the programming code. For example, if we apply the UML modeling language – the current de facto standard in software modeling – documents must be clearly organized and linked by means of roadmaps that lead from the use cases that describe the requested functionality to the technical diagrams and documents showing its implementation (class diagrams, reference to methods, components diagrams, etc.)

Linking the functional specifications with the technical design and with the repository of reusable and shared code will give rapid answers to the questions developers frequently ask themselves: what classes are involved in this
functionality? which method contains the code to validate the correctness of this process?

With this documentation, developers can be confident that they are making changes in the right place.

6. KM Project Implementation

The first step in the KM strategy is to recognize the importance of documentation, and the need to link documents created through the different steps in the software development process. These steps will change depending on the development methodology each organization applies.

The main target of the project we describe in the following sections is to reduce the time needed to develop a specific functionality by reusing programming code. This is an important source of benefits in those companies developing applications of the same type.

The second step in the proposed strategy is the creation of a repository of documents and source code. In this repository, source code will be linked to the functionality it implements, that is to say, to the documents and diagrams that describe this functionality.

To create the repository, we will complete these tasks:

Documentation is firstly created by means of UML diagrams and attached documents. UML modeling tools such as Rational Rose or TogetherSoft offer the possibility of generating documentation in html format. This documentation will contain all the diagrams and explanations included in the UML model. This documentation will be the basis of the subsequent steps.

Then, code is documented by developers, who will add a detailed description for each class, method, function, etc. they code. This comments are written inside the code. To distinguish the comments explaining the purpose and structure of the code from any other comments, a special markup is used. Developers should also add some metadata that will be used later to link the code to functional specifications or set of features.

By means of an automated process, source files containing code are parsed in a regular basis to extract the comments containing the metadata and documentation. For each class, operation and method processed, an html page is generated. This page will contain comments, metadata, and the source code itself. Metadata entered by developers are converted to html meta properties, to make possible their use in queries.

This process is similar to the JavaDoc tool, released by Sun Microsystems, to automatically extract comments and documentation from Java code and generate documentation in html format. A good explanation on how to accomplish this process and automate the extraction of comments in VB projects can be found in the current literature (Schultes, 2001).

The html pages are automatically moved to a repository where they are indexed. Indexing the full-text as well as the metadata properties gives the possibility of accessing the code used to implement specific processes or use cases, activities in a use case, product features and any business rule specified in the functional specifications.
Finally, the last step consists of linking the HTML pages generated from the code to the functional documents and diagrams generated in the first step. Pages are linked together by means of a navigable table of contents.

This table of contents contains references to HTML pages describing use cases, activities, sequence diagrams, classes and components. To update this table of contents with all the necessary links, an automated process was developed.

This process reads the metadata available in the HTML page that describes each piece of code and identifies the name of the class or method. As this information is also available in the UML models, it is not difficult to insert the links in the appropriate location.

As a result, developers can access an updated repository of source code linked to the functional documentation. This enables them to search for the implementation of specific functions or features. Full-text and metadata searching capabilities and the navigable table of contents help them locate the code and functions.

7. Additional features and next steps

The availability of a shared repository of code is the first step toward a knowledge management and reuse strategy. But sometimes, two programmers may code the same function in different ways. This may affect the quality of code, and errors can be propagated inadvertently.

To avoid this problem, a separate step that must be implemented is the revision of code by senior programmers. Senior programmers will be in charge of reviewing the code, fixing possible bugs, and identifying the best approach to solve a problem when two or more alternatives are available.

This is also a best-known practice in KM programs. The role of the knowledge guardian or subject expert who reviews and validates contributions to the knowledge base can also be applied in the software development industry. In fact, this is not something new, as the role of the senior programmer who reviews and validates the code written by other programmers has always been recognized as a must-have in different methodologies and software development practices.

The implementation of the validation process adds the requirement of dealing with two possible scenarios:

In the first one, code is published once it has been reviewed. So, we need to distinguish between validated and not-validated contributions. This ensures a higher quality of the code available in the repository, but makes knowledge transfer slower.

In the second scenario, code is moved into the repository automatically, with no prior validation. If any inconsistency is identified, the programmer in charge of maintaining the repository will have the possibility of rejecting a piece of code from the repository when a better alternative is already available. People who have reused this code should be automatically notified about this change.

Other functions that go beyond the current proposal are the possibilities of creating and managing discussion forums, adding comments or suggestions to code in the repository, or scoring the existing code. This will be an important issue to measure the effectiveness of the code, its reuse and the success of the whole implementation.
Other methods to measure the success of this KM program might consist of the counting of access to the repository or the LOC (lines of code) shared or taken from it.

Anyway, although any of these approaches might be used, we strongly recommend a different approach based on the scoring that developers assign to the implementation suggested by other colleagues, as the main purpose in the KM strategy is leading to valid solutions.

References
Framing Information

Abstract: The distinction between semiotic, semantic and ontological classifications is introduced. A few examples of semantic and ontological categories are then provided and discussed. The thesis is defended that semantic categories depend on ontological categories.

1. Introduction

One of the most striking features of recent developments in science and technology is the over-abundance of information. Unlike the societies of the past, modern ones are no longer afflicted by a lack of information. If anything they suffer from its excess, from having to cope with too much information. On the other hand, although the information may be stored somewhere, all too often one does not know where; and even when one is aware of how to find the information, it is often accompanied by further information irrelevant to one's purposes. And when information is available, it is often forthcoming in the wrong form, or else its meaning is not explicitly apparent.

However broad the range of information already gathered may be, a great deal more has still to be assembled and codified. And this inevitably complicates still further the problem of the functional, flexible, efficient and semantically transparent codification of information.

The events and objects of our experience are classified in many different ways: some forms of classification depend on the way in which events and objects are perceived (seen, felt, heard, smelt, tasted, together with the interactions among these perceptions); others depend on the way in which reality is conceptualized (e.g., in terms of its salient features); others depend on the ways in which perceptive and cognitive analyses are codified in linguistic expressions and structures; and yet others depend on general patterns or universals wholly intrinsic to the events and objects of the world. I distinguish three basic types of classification: the semiotic, the semantic, and the ontological.

Semiotic classifications distinguish between codes (linguistic, musical, pictorial, etc) and set the modeling environment for their proper analysis. This means that the usual series of types (string, floating, integer and boolean) should be extended to include at least, say, 'pixels' and 'voxels' (for digital 2D- and 3D-figural units), 'pcm-quanta' (for digital acoustic units; note: pcm = pulse code modulation), or 'phonems' (for phonological units), etc. The semiotic dimension therefore analyses the features of the signs used to convey information. Its main, but by no means only, sub-section is the language analysis module (semiotic aspects are further analyzed by Poli and Mazzola 2000).

Semantic classifications analyse forms of contextual and functional dependence. Moreover, they collect and seek to coordinate both de facto and
formally established standards. (I have reluctantly decided to resort to the overabused “semantic” owing to my inability to find a better term.)

Ontological classifications analyse the deepest internal (structural) features of the items under analysis. The paper is divided in two sections, plus this introduction and a few words of conclusion. Section 2 presents a few data on semantic classifications and Section 3 will analyse ontological classifications.

2. Semantic classifications

Besides semiotic distinctions, we also need a large quantity of semantic categories if we are to describe and classify items. The easiest point of departure for semantic categorization is the unity of the item. One then proceeds by classifying the item with respect to related criteria of analysis. Before I continue, note that the just mentioned point of departure for semantic classification depends on the ontological category of whole: an item has unity (i.e., is one) if and only if it is a whole. It is therefore evident that ontological categories underlie semantic categories. I shall return to this problem in Section 3.

Subsequent steps concern the categorization of (application of categories to) the item(s) under analysis. Here I shall concentrate on the widely accepted strategy which relies on the distinction between natural kinds and dependent kinds. In general, there is broad agreement that electron, proton, neutron, narcissus, chimpanzee, stickleback, carbon, gold and water are natural kinds, whereas table, nation, banknote, rubbish, cliff, perennial and bush are not. It is likewise generally agreed that, if there are natural kinds, they fall into at least two groups. There are kinds of stuff, such as carbon, gold, water, cellulose, and there are kinds of object, such as tiger, chimpanzee, stickleback, narcissus (Wilkerson 1995). As before, stuff and object are ontological categories. Besides the trivial cases of the categories stuff and object, there is nothing to prevent acceptance of other ontological categories, like those of process and group (and perhaps even further less acknowledged categories: for details see Poli 2001).

Dependent kinds may be further subdivided into functional kinds, such as table and banknote, and contextual kinds, such as cliff and bush.

Crystal-clear cases aside, it may be difficult to properly distinguish cases of natural, functional and contextual items. The following may be one possible (partial) way out of the problem for the natural/functional opposition. Let us assume that functional kinds are linked to technologies, that is to say, to practices employed to name, recognize, use, produce and modify items of various types. The argument that I wish to develop is that the relation which connects natural and functional kinds is of the same nature as the relation between sciences and technologies. For this reason, it is advisable to dwell for a moment on the difference between sciences and technologies. The key reason for linking natural kinds to the sciences is that the latter, however much they are obviously interconnected, represent irreducibly different points of view. “An excellent reason for taking biology seriously is that the biological properties of things obviously depend directly on their physical and chemical properties. But the explanatory apparatus of biology cannot in practice be reduced to the explanatory apparatus of physics or chemistry or both” (Wilkerson, 1995, p. 39).

In other words, the explanatory apparatus of sciences is emergent. Any science has its own ‘window’ on the world, selecting (in its own way) only those
objects that are at the ‘right’ level of magnitude, energy and complexity. On the other hand, the explanatory apparatus of technologies is not emergent. “It is, as it were, ... constantly being reduced to, or connected with, or supplanted by, the explanatory apparatus of some discipline characteristically concerned with entities of some lower level, notably physics, chemistry and biology” (Wilkerson, 1995, p. 40).

Therefore, we may say that items described by sciences are items of a certain natural kind, whereas items described by technologies are items of a functional kind. Obviously, there is no clear-cut distinction between sciences and technologies: science is involved in numerous technologies, and technologies use the most disparate of sciences. Even if the proposal just put forward may not be able to settle all the possible cases, it nevertheless provides some guidelines.

The other remaining case concerns contextual kinds. The latter can be well represented by the following examples:

Gardeners talk cheerfully of seedlings, saplings, trees, shrubs, bushes, climbers, perennials, annuals, pot plants, and so on, but none of these terms pick out a real essence; none are likely to appear in reports of serious scientific investigation; and none refer to a kind determined by an intrinsic property. One and the same plant will grow as a tree under one set of conditions and as a shrub under others (e.g. many Eucalyptus and Acer species). One and the same plant will be an annual or pot plant in a temperate European climate and a shrub in a hot African climate (e.g. Pelargonium species). One and the same plant is a shrub in western Ireland and a hardy perennial in Nottingham (e.g. Fuchsia magellanica). None of those terms pick out an intrinsic property and none of them correspond, even approximately, to any botanical classification ... Yet none of the terms has any connection with convention, artifice or culture (Wilkerson, 1995, p. 37).

The same point can be made about geographical and meteorological kinds:

Geographers talk of beaches, cliffs, mountains, valleys, seas and volcanoes. Meteorologists talk of depressions, anti-cyclones, winds, thunderstorms, clouds and hurricanes. But the terms do not pick out things with real essences, they do not figure in scientific generalisations and they do not pick out any relevant intrinsic properties. One and the same lump of material will count as a mountain in one environment, as a valley floor in another, and as part of the sea bed in yet another (Wilkerson, 1995, pp. 37-8).

The main feature of contextual kinds is that they contain an essential link with their environment. If you change the environment, contextual items may present different (even radically different) features.

Further developments concern hybrids between natural and dependent categories or between functional and contextual ones (for lack of space I cannot dwell any further on these: for details see Wilkerson, 1995 and Poli, 2001).

Before concluding this section, the problem of the many dependence relations between kinds should be briefly considered. Consider for instance the connection between the natural kind wood and the functional kind table. From the point of view of categorial analysis, what are the features of the complex whole table made of wood?

Put otherwise, what information is conveyed by its natural component (wood) and what information is conveyed by its functional component (table)? It takes a little thought to see that in this regard we are not too distant from Aristotle. If we know that the object over there is a table, we can make safe predictions about its behaviour in certain circumstances. In fact, we know the likely outcome of putting the kitchen table on a bonfire (Wilkerson, 1995, p. 34). The point to the
stressed is that, in making our predictions, we are exploiting the fact that every item belongs to at least one natural kind or is dependent on an item belonging to at least one natural kind.

In its turn, the theory of dependence referred to here requires a theory of levels of reality (see section 3 below). Let us suppose that we have developed one. Thus the general thesis becomes: for each of their ontological levels, all items belong to at least one natural kind and to one or more dependent kinds (functional or contextual). For any level, the natural kinds of every item are connected with the causal links of the item (as the above example of the table on the bonfire indicates).

This section has presented a few of the intricacies lurking behind a fragment of the field of semantic classifications, the one that relies on the distinction between natural and dependent kinds. The two main results of such analysis is that (1) semantic categories rely on ontological categories, and (2) items having a pluristratified nature are such that at least some of the subitems pertaining to one of their levels belong to a natural kind. (Other forms of semantic classification besides the opposition between natural kinds and dependent kinds are discussed in Poli, 2000, 2001.)

3. Ontological Classifications

Of the various types of classification, the ontological one is the most difficult, for a number of reasons (some of which have been analysed by Poli, 2001, 2002). In this section I shall try to give a very simplified idea of the highly complex structure of ontology.

The first step is to distinguish between what we are talking about and its determinations. Resorting to traditional terminology, I shall address the first topic as the problem of "substance", and the second one as the problem of "determinations". For the time being, I shall only consider the problem of substance.

My basic tenet is that the theory of substance comprises at least three sub-theories: the theory of particulars, the theory of levels of reality, and the theory of wholes and their parts (all of them alluded to in the previous section). Most traditional and contemporary theories of substance fail precisely because they lack one or more of the above sub-theories.

Let us continue to use "item" as the most generic descriptive term. Subsequent distinctions should consider at least objects, processes, stuffs and groups. Each of these has its ontological features and deserves its own theory. Higher-order items are items composed of other items (groups of processes, as for parallel computing, etc). (Poli 2001, distinguishes eight different types of basic item.)

As to the problem of levels, we may distinguish at least three ontological strata of the real world: the material, the psychological and the social. Specific forms of categorial and existential dependence exist among these strata. For example, a psychological item or event requires an animate physical object as its existential bearer. Should there be no person (and should there be no body of some such person), then neither will there be the correlative psychological states.

A relationship of matter and form holds among many items. In these cases, matter and form are correlative categories, so that any form may be the matter of a higher form, and any matter may be the form of a lower matter. The hierarchy thus constituted is a progressive overforming of matter and form. The nature of the
physical world is clearly governed by this embedding principle: the atom is the matter of the molecule, but it is already an entity endowed with form; the molecule is the matter of the cell; the cell is the matter of the multi-cellular organism; and so on.

However, not all the dependences that structure the world are of a matter/form type. When one moves from the organic to the mental plane, one finds a dependence relation that is not reducible to the matter/form relation. One cannot say, in fact, that atoms or cells or organisms are the matter of the mind. Organic reality takes atoms and molecules and assembles them into a new form, consciousness, which is nevertheless not made up of organic forms. In the passage from the material to the mental there arises a new series of forms whereby corporeal life with its forms and processes no longer functions as matter. The organic levels are mirrored in psychic life: they influence it, they follow close upon it, but they are not part of it. In effect, the life of the mind does not comprise organic processes, nor does it use them as its building blocks, even though it is supported by them and is influenced by them.

One finds another break in the social stratum. In both these cases the dependence relationship is no longer of matter/form type but becomes one of a completely different kind: a bearer/borne relationship. In these cases, the substratum of the higher level is not the matter of the lower level (Hartmann, 1952, pp. 68-69).

Belonging to the social stratum are all phenomena of communication, and therefore the complex of social phenomena and customs, economic and legal realities, history, language, science, technology and the body of knowledge of every epoch, and morals.

Description of the strata and levels of reality intersects with description of the items of which it is composed. We humans participate in all three strata (although we do not exhaust the multiplicity of any of them). We have a material (organic) base, we have a mind, and we are simultaneously social beings. But our material base is one of the many material bases offered by the natural world; just as our mind is only one possible mind, and our participation in the social world is never such that we can absorb it in all its aspects (Poli, 1998, p. 2001).

The third dimension of the problem of the substance comprises the theory of wholes and their parts. Without entering into detail, three kinds of wholes at least may be distinguished: aggregates, wholes in the proper sense and systems. Aggregates consist of proximate parts. Wholes in the proper sense comprise parts "which go together". Systems require a dynamic exchange between the whole and its parts.

Unity by solidarity is stronger than unity by proximity. This means that only some items that are aggregates are also (integral) wholes, and that some of the latter may be systems. Needless to say, the most difficult task is furnishing an adequate characterization of the dynamic components of systems. For the moment I merely point out that it is possible to determine various forms of dynamic unity, ranging from those that obtain in material systems of a physical nature to those that obtain in systems which, like living and social systems, are able to produce the elements of which they are composed.

The above classification can be further clarified by adding that aggregates are characterized by relations among their parts. Wholes require both part-part relations and parts-whole relations, whereas a proper characterization of systems
requires information on three kinds of relations: part-part relations, part-whole relations and whole-part relations (Poli, 2001, chp. 7).

Even if the above is only a very rough sketch of the basic structure of an ontology, it may nevertheless provide the reader with initial understanding of its main tenets. The point is relevant because, for some time, the term and idea of ontology have begun to enjoy currency in various sectors of the information processing community (and particularly in groups working in the fields of (i) knowledge representation, (ii) databases, (iii) natural language processing, and (iv) automatic translation). In short, those who most frequently talk about ontology are researchers in the acquisition, integration, sharing and re-utilization of knowledge (Guarino and Poli, 1995, Poli, 2001).

Ontology comes into play as a viable strategy with which, for example, to construct robust domain models. An ontologically grounded knowledge of the objects of the domain should make their codification simpler, more transparent and more natural. Indeed, ontology can give greater robustness to models by furnishing criteria and categories with which to organize and construct them; and it is also able to provide contexts in which different models can be embedded and recategorized to acquire greater reciprocal transparency (Poli, 1996; Poli and Mazzola, 2000; Poli, 2001).

Traditional philosophical ontology (like the one sketched above) and its new understanding (call it ontology as a technology) are still very distant from each other. This comes to no surprise: ontology as a technology is still in its early stages of development. At the moment, the artificial intelligence research community seems to have reached broad agreement only on the need for formal standards. This is indubitably an important step forward, but it is one which we can call authentically ontological only because of its improper extension of the concept of ontology (understood in its strict sense as categorical analysis; on this more usual sense of ontology see Poli, 1992, Poli and Simons, 1996).

An example may be of help. It is well known that important standards for software construction have become established in recent years. Suffice it to mention the Standard Template Library (STL) for C++. This is certainly a positive development, but it is one that concerns solely formal components, not ontological ones. We will be able to talk of a similar development in ontology when we have a Standard Template Library for ontological patterns, like PROCESS, THING, EVENT, PART and WHOLE (Poli, 2001).

We have seen that ontology has a complex conceptual structure. Without any awareness of its complexity, it is not possible to devise the most appropriate methods of analysis. On the other hand, it is the world that is complex, and there is no reason why it should be amenable to hypersimplified codification.

4. Conclusion

The paper has defended the theses that there are at least three main different types of classifications (semiotic, semantic and ontological), and that at least some semantic classifications require and rely on underlying ontological categories. Unfortunately, most contemporary scholars seem unaware of the challenging complexity of the ontological framework and this may undermine the value of even major projects.
References


Structured Models of Scientific Concepts for Organizing, Accessing, and Using Learning Materials

Abstract: The knowledge represented in learning materials for the sciences is typically organized around term-based or "weakly-structured" models of concepts and their interrelationships. We introduce a "strongly-structured" model of scientific concepts that provides the foundation for a knowledge base (KB) of concept representations. An extension of the Alexandria Digital Library employs such a KB, together with associated collection and services, to support undergraduate learning.

1. Introduction

Concepts and their interrelationships are the fundamental building blocks for representing the phenomena investigated in mathematics, science, and engineering (MSE). The value of any MSE concept rests heavily on the degree to which it: (1) is objectively representable; (2) possesses well-defined semantics; and (3) is informative. These attributes permit scientists to communicate unambiguously about how to represent, use, and interpret scientific concepts and to derive useful information from representations of phenomena that are based on the concepts.

It follows that an understanding of phenomena within the context of MSE depends on an understanding of the concepts used in building representations of the phenomena. Learning environments for MSE should reflect this principle in explicit ways. Towards this end, we describe digital library (DL) services that support the process of learning in the domains of MSE. These services provide access to, and use of, learning materials in terms of strongly-structured models of scientific concepts and their interrelationships.

In this paper we briefly discuss the term-based models of concepts that are typically employed in learning materials and libraries. We then introduce the notions of "strongly-structured" models of concepts and of a knowledge base (KB) of strongly-structured representations of concepts and their interrelationships. Such a KB, together with associated DL collections and services, forms an extension of the Alexandria Digital Library (ADL) environment that is intended to support undergraduate learning.
2. Weakly-structured models of concepts

In many text-based learning materials, concepts and their interrelationships are primarily denoted by linguistic terms or alphanumeric symbols. This is reflected in the widespread use in textbooks of glossaries of terms and symbols. Many of the important and defining characteristics of MSE concepts, such as their representation, semantics, properties, relationships to other concepts, and use cannot be represented in such simple linguistic terms, as in the case of structural representations of complex chemical compounds. Instead, such knowledge is typically distributed in unstructured ways throughout the learning materials. Many important aspects of MSE concepts are not organized in terms of any explicit knowledge representation schema (KRS). We may therefore characterize such term-based representations as weakly-structured models of concepts.

Term-based representations are also the primary mode for representing concepts in KRS that support the access technologies of traditional and digital libraries. Examples of such KRS include: (1) term lists, such as authority files, glossaries, gazetteers, and dictionaries; (2) classification and categorization schemes, such as subject headings, hierarchies, taxonomies, paradigms, faceted analysis, and categorization schemes; and (3) relationship groups such as thesauri, semantic networks, frames, ontologies, concept maps, and topic maps.

Weakly-structured representations of concepts have value in learning environments. They support, for example, access to traditional knowledge containers, such as texts and journals, in which term-based representations of concepts occur. They are also of value in supporting high-level graphical views (or "concept maps") of the interrelationships among concepts. On the other hand, weakly-structured representations of concepts cannot easily support access to, or integration of, knowledge concerning many of the attributes of concepts that make them useful in MSE modeling activities.

3. Strongly-structured representations of concepts

The inadequacies of weakly-structured models of concepts for many learning and research applications are made apparent by the efforts of various MSE disciplines to develop models of MSE concepts that represent their key attributes in explicit form. These models focus on such attributes as the objective representations, operational semantics, use, and interrelationships of concepts, all of which play important roles in constructing representations of phenomena that further understanding of MSE domains of knowledge. We characterize such representations as "strongly-structured" models of concepts.

The MatML Working Group of the National Institute of Standards and Technology (NIST) has, for example, constructed a model for representing concepts relating to Materials Science (NIST, 2001.) In particular, they have created an abstract DTD for representing such concepts as XML records. At the highest level, the basic model for representing materials concepts takes the form: 

```xml
ELEMENT Material (BulkDetails, ComponentDetails?, Terms?, Graphs?)
```

in which 

- `BulkDetails` contains a description of the bulk material and
- `ComponentDetails` contains a description of the components comprising the bulk material.

`ComponentDetails` can be used, for example, in describing the base metal, the heat affected zone, and the weld metal of a welded material.
As a second example, the Chemical Abstracts Service (CAS) Registry (Weisgerber, 1997) employs a strongly-structured model of chemical substances that includes systematic chemical names, their various representations (molecular structure diagrams, molecular formula, index of ring systems), and information about the properties and interrelations of substances.

4. A strongly-structured model of SME concepts

Based on these examples and previous work (Smith et. al., 1995) we have developed a strongly-structured model of concepts for SME domains in terms of a frame-based KRS with slots and attribute-value fillers. The model, whose framework is shown in Figure 1, is implemented as an XML schema. This schema is used as the basis for creating domain-specific KBs containing XML records of concepts. Intended applications of the model are in learning environments.

We briefly discuss salient aspects of the schema and illustrate it in terms of the (mathematical) concept of a plane polygon (PP).

- **PreferredTerm** may be interpreted as the “usual” term-based representation of a concept. For a PP, “n-gon” may be viewed as a non-preferred (alternative) term.

- **Description(s)** (or definitions) of a concept, represented as free text, are intended to be high-level, descriptions of the concept from one or more authorities. For example, a PP may be described as “a plane geometric figure with a simple boundary composed of n connected, straight-line edges.”

- **HistoricalOrigin(s)** is a free text element providing background on the evolution of the concept.

- **Relationships** between a given concept and other concepts in the KB constitute the most important element of the concept model.

- **KnowledgeDomain** relationships indicate the academic disciplines and/or topics, themselves represented as concepts, in which a given concept finds use. Euclidean Plane Geometry, for example, is a topic within the discipline of Mathematics in which the concept of a PP finds application.

- **ScientificUse(s)** relationships indicate applications of a concept (using Applications) and roles within the structure of scientific activity (using Class) with applications and roles themselves being represented by appropriate concepts. An application of the concept of PP, for example, is in approximating complex geometric figures in the plane.

- **Representation** relationships of a concept define how it may be represented in terms of other concepts and includes mathematical representations of the relationship. It is, perhaps, the single most important element in the concept model since the value of an MSE concept depends on the degree to which its representations may be manipulated in order to infer information of interest concerning some phenomenon. We may, for example, represent a polygon as an ordered list of n two-dimensional points \( <P_1, \ldots, P_n> \), from which we may compute the area of the polygon and other useful attributes.
We represent the relationship in both directions (HasRepresentation and PartiallyRepresents) and differentiate between explicit full, explicit partial, implicit full, and implicit partial representations of a concept. The term “implicit” indicates that additional information processing must be applied to a representation in order to obtain explicit representations that can be manipulated directly for purposes of inference. An equation, for example, contains an implicit representation of the concept corresponding to its solution. A PP (without holes) can be represented: (1) explicitly and fully as a tuple of two-dimensional points; (2) explicitly and partially as a sequence of a proper subset of such vertex points; (3) implicitly and fully as the intersection of a set of two-dimensional half-planes; and (4) implicitly and partially as a distorted graphical representation of a polygon from which explicit representations can be obtained by estimating the locations of the vertex points.

- **DefiningOperation(s)** provide the operational semantics for a concept. In the case of a PP, for example, the semantics may be partially defined in terms of how a particular representation of a polygon may be symbolically (or computationally) manipulated. The semantics of a measurable scientific concept, such as energy, may be defined in terms of appropriate measurement operations, themselves represented by concepts.

- **Property** relationships refer to other concepts that partially characterize a concept and are inferable from representations of the concept. Examples of properties for the case of PP are area, perimeter, and centroid, together with descriptions of procedures for deriving them from representations of a PP. These relationships are represented in both directions (HasProperty and PropertyOf).
Hierarchical relations, (SetMembership and Partitive), are two important relationships underlying thesauri. They are represented in both directions. For example, the concepts of triangle, quadrilateral, and pentagon form subclasses of PP while connected, closed, bounded, two-dimensional pointsets forms a superclass.

Causal relations refer to other concepts that are causally-related to the concept of interest, in the accepted scientific sense. They may be viewed in terms of Causes and Causedby. Causal relations are not applicable to the concept of PP.

Co-relationships indicate relationships between scientific concepts that are neither causal nor characterizable in terms of the property relationship. Among the properties of a PP, for example, it is clear that area and perimeter are co-related, but not causally related, and that they should not be viewed as properties of each other.

Examples (or instances) of concepts, together with pointers to their representations, are for illustrative purposes.

We note that the MatML and CAS models may be mapped into this general model of a concept (see, for example, Smith et.al., 2002.)

As noted above, the ScientificUse(s) relationship contains a Class sub-element indicating the concept's role within scientific activities. For pedagogical purposes, we have constructed a use-based classification of scientific concepts based informally on a National Research Council publication on scientific education (NRC, 1996). Our classification (Smith et.al., 2002) provides a basis for characterizing concepts in terms of objective scientific operations (i.e., the operational semantics of the concepts.) For example, the abstract mathematical concept of an arithmetic equation may be defined in part by objective syntactic operations that may be applied to symbolic representations of an equation; a measurable concept, such as stream velocity, may be defined in terms of objective operations used to determine a stream's velocity; while a methodological concept, such as hypothesis testing, may be defined in terms of sets of objective operations.

5. Concept KBs and their application in DL environments

The Alexandria Digital Library (ADL) Digital Earth Testbed system (http://www.alexandria.ucsb.edu) has been extended with: (1) a KB of scientific concepts, from the domain of physical geography, that are represented in terms of our XML schema for concept representation; (2) a collection of heterogeneous learning materials exemplifying the concepts and their properties in various contexts; and (3) services that provide a variety of views of the content of the KB and associated collection.

This extension to ADL is being deployed in teaching an introductory course in physical geography to about 200 students in Fall, 2002. Its KB of concepts, instantiated as a database of XML records supported by the Tamino system (http://www.softwareag.com/tamino/) is constructed from a variety of sources. These include, for example, the knowledge of domain experts and materials from various printed and electronic sources, including glossaries from the textbook of choice2.
The collection of materials is represented as a standard collection of ADL (Janee and Frew, 2002), which implies an ADL-searchable collection of metadata records. We have used an extension of the Alexandria-DLESE-NASA (ADN) joint metadata content model for cataloging such items (http://www.dlese.org/Metadata.) The extension involves a field for representing concept terms and elements. The collection is populated with a variety of materials, including the figures and tables from the textbook and other digital resources, such as maps and images from the operational ADL.

Various services support the use of the KB and collection in learning contexts.

- Services for creating, modifying, and deleting entries in the concept KB and the (metadata) collection of illustrative items.
- Tamino-based services for searching the concept KB and ADL middleware services for searching the collection of illustrative items.
- Services for creating textual and graphical views of the KB and collection. Important classes of views include:
  - Textual views of the concept KB (i.e., “subset” thesauri) that may be extracted from any subset of concepts and the various interrelationships between concepts. If, for example, the subset of interrelationships includes only the hierarchical relationships between concepts then textual views correspond to the heart of a traditional thesaurus.
  - Graphical views of the KB include visual representations of individual concepts or subsets of concepts and interrelationships, such as “all concepts causally related to the concept of soil-moisture content”. Such views may be usefully represented as three-dimensional lattices of elements and edges.
  - Graphical summary views of items in the collection. One may, for example, construct maps summarizing the geospatial locations of collection items representing a given concept element.
- Services for creating personalized collections (PCs) of items from the KB and the collection. An important example of a PC is a sequenced set of KB and collection items that may be used by an instructor in teaching a course, either for lecture presentation or for self-guided learning. We may interpret such a sequence as a trajectory through the KB of concepts, together with associated illustrative materials. Such trajectories may be viewed as the “conceptual backbone” of a course. These services support the re-use of KB and collection materials in creating learning materials.

6. Pedagogic use of the KB, collection, and services

Our applications of KBs of strongly-structured concept representations in learning environments build on recent studies in educational psychology (see Mayer, Smith, and Borgman, 2002.) These studies indicate that concept maps, defined in this context as “all terms used in a lesson with labeled links among them”, play significant roles in promoting student learning and assessing student knowledge. Educational research relating to rhetorical structures, or the ways in which materials can be organized into coherent structures for learning, suggests how the KB and collection may be used in furthering student understanding of some
domain of knowledge. In process rhetorical structures, for example, cause-and-effect systems can be represented by flow charts and maps. This corresponds to a view of the KB that extracts Causal and Hierarchical Partitive relations between a given set of concepts. Other rhetorical structures associated with different views of the KB include *problem solution* (problem based views), *classification* (hierarchical views), and *compare* (matrix-based views in which the attributes of different concepts are compared.) Recent research (see, for example, Mayer, 2002) indicates that greater depths of learning occur when rhetorical structures are used to organize learning materials.

Services supporting access to, and use of, the resources in the KB and associated collection enable the construction of views corresponding both to different rhetorical structures and the personal preferences of instructors. Scenarios exemplifying the pedagogic use of such KBs, collections, and services may be found in Smith et.al., 2002.

7. Conclusions

It will take time to assess the pedagogic value of using KBs of strongly-structured representation of SME concepts, together with the associated collections and services, in learning environments. Recent developments, in both text publishing and educational psychology indicate, however, their potential value. Furthermore, the development or KBs of strongly structured concept representations by such organizations as NIST's MatML Working Group or the Chemical Abstracts Service (CAS) suggests the potential value of constructing such KBs as resources for students and researchers in many domains of SME knowledge.

Acknowledgments: The work described herein was partially supported by the NSF-DARPA-NASA Digital Libraries Initiative and the NSF NSDL initiative under NSF IR94-11330, NSF IIS-9817432, DUE-0121578, and UCAR S02-36644.

Notes
1 Members of the ADEPT Knowledge Organization Team: Olga Agapova, Olha Buchel, Michael Freeston, Jim Frew, Linda Hill, Richard Mayer, Jian Qin, Laura Smart, Tim Tierney, Alex Ushakov.
2 Elements of Geosystems by R.W.Chistopherson and published by Prentice Hall, who have generously provided ADEPT with electronic versions of the complete book.

References


A KR Multi-Hierarchies/Multi-Views Model for the Development of Complex Systems

Abstract: Recent methodologies in knowledge representation are oriented towards the construction of multiple representation models. The concepts commonly used in describing levels of granularity of knowledge are based on conceptual and symbolic levels. We propose to describe the conceptual level of a KR system by introducing multi-views and hierarchical levels which allow to represent explicitly a complex KBS.

1. Introduction
The resolution of complex technical problems as design, diagnostic, planning or process control requires separate phases for reasoning (Davis, 1982) (Genesereth, 1982). In each phase, for the same modeling domain, in general different models are necessary (Booch, 1997). Furthermore, for each model, one or more conceptual views are associated. A view has for goal to reason on concepts and aspects relating to a certain stage. However, a conceptual view, in its complexity, can improve for its representation, a hierarchical approach that leads to different abstract levels or comprehension levels. Therefore, we present in this paper a knowledge representation Multi-Hierarchies/Multi-Views (MHMV) Model which allows complex systems to be tackled in this way.

The remainder of this paper is organized as follows: Section 2 describes the notions of the KR MHMV model. Section 3 presents the basic concepts for representing the MHMV model and details the MHMV link concepts and its reuse. Finally, Section 4 concludes this article.

2. The KR multi-hierarchical/multi-views model
In the proposed KR MHMV model a complex system is described according to the following concepts:

2.1. Model
A model is a representation of a given system with respect to a given set of problems. In general, to two different classes of problems will correspond, for the same system, two different models. So each class of problems leads to a specific domain of modeling. So we consider that a model is a simplified representation of the structure and behavior of a real system designed to support and facilitate reasoning concerning the system.

For example, a digital circuit system can be represented by a digital-behavior model for function-related applications such as simulation and by a thermal model for heat dissipation applications.

2.2. Views
During a particular application in a particular domain of modeling, a number of conceptual views of the modeled system may be required. During the design
process of a digital system, for example, three separate views are considered by the

Due to the complexity of the views, it is often necessary to decompose a
view at different hierarchical levels of comprehension. Defining a suitable
hierarchical level reduces the amount of information to be considered to
manageable proportions.

However, this decomposition of a view of a model (Gribble, 1998) (Wang,
2000) in hierarchical levels is not necessarily isomorphic to the decomposition of
others views of the same model.

2.3. Hierarchical levels

In the KR MHMV model we have identified three kinds of hierarchy in a
same view:

a- The abstraction hierarchy in a view: represents an appropriate
comprehension level for the view decomposition. An abstraction level permits to
limit the content and the sense of information to consider. Details that are not
pertinent are relegated at a lower abstraction level. The abstraction level of a view
includes:

- the data structure model
- the tools used by the application builder to specify the reasoning type at this
level

b- The description hierarchy: enables to have a finest description of a same
behavioral element belonging to a given abstraction level by keeping the same
specifications of its inputs and outputs and decomposing it according to different
sub-structures. So, each abstraction level is composed of one or several description
levels. This hierarchy has a meaning only in behavioral views.

c- The conceptual hierarchy: permits to organize the concepts by generality
levels. Each description level is composed of one or several conceptual levels and is
associated to it a library based on a conceptual hierarchy. This library constitutes
the base of reusable components.

3. MHMV model representation

In the KR MHMV model, we have extended the graph formalism in such a
way that hierarchies, views, links can be clearly defined. The extension of the graph
formalism consists of the following concepts:

3.1 Basic concepts

- Node: is a unit of description of a basic element or a function belonging to
a given hierarchical level of a model view. A node component is composed of a
connector and the description of its structure and/or its behavior.

- Graph: is a representation, based on a logical or physical structure of the
modeled system, which consists to inter-connect elementary node components
considered like that with regard to the level of abstraction where the system is
represented. A graph component is composed of a connector, nodes, links, and
transfer links allowing the information transferability from the graph connector to
its nodes connectors and vice versa.

. To each node can be associated a sub-graph which represents this same
node at a lower abstraction or description level of hierarchy.
- **Link**: considered as a component, is composed by a *source connector*, a *destination connector*, *semantic links* (association, composition, inheritance,...) and *transfer links* which give information about the transformation of information leading by the link.

- **Connector**: is associated with graphs, nodes and links components and is composed of ports. It represents a well-defined interface between a graph, a node, a link component and the surrounding environment.

- **Port**: is an entity attached to a node/graph/link connector and is the only means by which data can enter or leave nodes, graphs or links components.

### 3.2 Link component in KR MHMV model

A *link component*, in addition of its connectors, is specified by semantic links (the description of the relation types between the two linked components) and transfer links (the transferability of information between components).

a- **Semantic links** allow the definition of semantics for inter-component references. In many models, the composition link, for example, conveys strong semantics (Rumbaugh, 1988) (Kim, 1989): a composition link can, for example, be reflexive, transitive, exclusive or shared, dependent or independent, predominant or not. We think that all this information may be included in the reified concept of semantic link and then be specialized according to different kinds of semantic links.

b- **Transfer links** are carried out using their connectors. They may be explicitly established by:

- defining correspondences between ports of their source and destination connectors,
- attaching appropriate translation functions to these correspondences.

### 3.3 Reusing transfer links for different semantic links and as adapters

A question that springs to mind is: why separate transfer link from semantic link?. The answer is that it enhances reusability as the same transfer link can serve two or several different semantic links. The idea is to separate the semantic properties (structural, behavioral) of a link from its operating mechanism (information transferability, information propagation). For example, the *derivation* link used for evolution management in object design permits linking versions of a component to the initial one. A version can be derived by deletion of, addition to, or modification of an existing component. Many methods have confused derivation link with inheritance (since inheritance permits addition and modification). Some alter the inheritance link so that it fits derivation, using inheritance with exceptions or adding constraints. We can reuse the transfer applied to inheritance, adapt it so that it can support information-deletion and reuse it for this link.

Otherwise, the transfer link can also serve as an *adapter* between two MHMV components with the same or different types among nodes, graphs and links components exactly like a digital component adapter.

We consider in the KR MHMV approach that the node/graph/link components must be defined in free-context, according to an abstraction level or description level and constitute the fixed part of a potential reuse and adapted to a given context (in-context) thanks to a transfer link (adapter) which constitutes the variable part of a reuse.

### 3.4 Links Classes in KR MHMV model

Several kinds of links occur in the KR multi-hierarchical/multi-views model definition:
Inter-levels link allow passage from abstraction, description or conceptual level to inferior or superior one. In inter-level links, we have identified general links as inter-abstraction levels, inter-description levels and inter-conceptual levels links.

For inter-abstraction levels links, we distinguish, for instance, the expansion links which allow transformation of input/output information node/graph to the lower abstraction level. Compression links represent the opposite links compared with expansion links. The inter-abstraction levels links are based on the semantic links of UML association links (Muller, 1997) with mapping and arithmetic transfer links.

Among inter-description levels links, we can mention composition links which represent the existing links between a node belonging to a description level \( N \) and nodes corresponding to the same element at the lower adjacent level. The aggregation link represents the opposite link. The inter-description levels links are based on the semantic links of UML composition and aggregation links (Muller, 1997) with identity transfer links.

Inter-conceptual levels links are based on the semantic links of UML inheritance links with identity mapping transfer links. Thanks to the explicit description of a transfer link, we can build for example different kinds of inheritance: strict inheritance, inheritance with exception, inheritance with renaming, ...

Intra-level links: They are existing links between two nodes belonging to the same conceptual level.

Inter-views links: They are composed of change views links and exchange of information links between two views. In the transformation of a view in another one, we can distinguish two cases:

a- the transformation is established between two descriptions of the same level of abstraction belonging to two different views. This transformation is generally algorithmic. For example the passage from an algebraic description of a boolean function to a logical gate representation.

b- the transformation is established between two different levels of abstraction of two models views. This transformation is generally based on inter-abstraction levels links and heuristics. For example, the passage from the physical view to the structural one.

The different KR MHMV links presented before can be obviously instantiated, specialized and completed by the user according to his needs.

To tackle the use/reuse problem of models, we have defined the MHMV link component as a composite object composed of semantic and transfer links which allow to specify separately different semantics relationships and information transfers between components. This description is made by an external and independent way from the connected components (free-context). Therefore, we have defined, an ontology for semantic and transfer links (Oussalah, 1999). This ontology is used in the KR MHMV model. One of the most important contribution in KR MHMV model is to be able to describe in an independent way all views and abstraction levels of a model and to link them via the different KR MHMV link components.

In our basic framework, we proposed two kinds of predefined semantic links: inclusion and association, and four kinds of transfer links: arithmetic, mapping, user and logic links. These different links can then be completed according to the application needs.
4. Conclusion

We have defined a KR model for developing complex systems. The KR MHMV concepts defined are generic and can be applied to different types of systems. We believe that the KR MHMV model can contribute to the creation of better quality solutions in the area of complex systems design.

The main contribution of our work is, firstly, the explicit description it offers of the different KR MHMV concepts; second, the possibility to extend and adapt these concepts; and finally the natural use of the graph formalism for representing them: there is no gap between the defined concepts and their implementation.

References


A Unifying Model of Document Relatedness for Hybrid Search Engines

Abstract: Previous work on search-engine design has indicated that information-seekers may benefit from being given the opportunity to exploit multiple sources of evidence of document relatedness. Few existing systems, however, give users more than minimal control over the selections that may be made among methods of exploitation. By applying the methods of "document network analysis" (DNA), a unifying, graph-theoretic model of content-, collaboration-, and context-based systems (CCC) may be developed in which the nature of the similarities between types of document relatedness and document ranking are clarified. The usefulness of the approach to system design suggested by this model may be tested by constructing and evaluating a prototype system (UCXtra) that allows searchers to maintain control over the multiple ways in which document collections may be ranked and re-ranked.

1. Document Network Analysis, Graph Theory, and Matrix Algebra

Within the broad discipline of information studies, it is becoming increasingly useful to delineate and refer to a particular confluence of research questions by applying a new, encompassing label: document network analysis. The subfield denoted by this term may be conceived as originating at the intersection of four existing specializations: (a) information retrieval (IR) systems analysis and design, (b) bibliographic (library) classification, a.k.a. knowledge organization, (c) hypertext, and (d) bibliometrics. A central, common assumption made by scholars working in each of these areas is that the primary objects of study — documents in which human thought is recorded, and through which representations such thought is communicated — are objects that are created, organized, transferred, and used on a mutually dependent basis. No individual document exists independently of all others; every document is related to every other, in various ways, and to extents that vary both with time and with the identity of the observer. The universe of documents may be viewed as a vast, dynamic network in which individual documents are situated with regard to one another through specification of the relationships that are perceived to exist among them at given times. In summary, the identity and unity of document network analysis are derivative of the following shared viewpoints of its practitioners: (a) an epistemological stance that accepts the reality of causal connection between structure and process, and a corresponding conviction that theory about the latter may be tested by examination of the former; (b) an ontological perspective that privileges the relational properties of objects; and (c) a methodological preference for mathematical methods of structural analysis.

Models of document networks typically make use of concepts drawn from the branch of mathematics known as graph theory (Harary, Norman, & Cartwright, 1965), which may be used to model any structure consisting of a set of objects...
(vertices, or nodes) and a set of relationships (edges, or links) among those objects (Dipert, 1997). The use of mathematical techniques of graph theory and matrix algebra in the modeling of document networks has been well documented over several decades, and often mirrors the prior application of such methods in related fields such as chemical structure analysis and social network analysis. Graph theory has been explicitly applied in the analysis of hypertext structures (Furner, Ellis, & Willett, 1996) and citation structures (Gamer, 1967). Similarly, matrix algebra has long been a core component of methods of automatic classification (clustering) of documents (Van Rijsbergen, 1979), and of citation analysis (Yagi, 1965). More recently, it has been put to work in the identification of “authoritative” websites (Kleinberg, 1999). At an even greater level of generality, basic vector-processing techniques involving the measurement of the degree of similarity between single vectors lie at the heart of content-based (Fairthorne, 1956), collaboration-based (Resnick, Iacovou, Suchak, Bergstrom, & Riedl, 1994), and context-based (Salton, 1963) retrieval systems.

2. Retrieval as Ranking

A central tenet of document network analysis, then, is that every document is related to every other. But what varies is not only the kind, but the strength of relationships — i.e., the degree of relatedness, the extent to which a particular pair of objects are related in a particular way. Objects may be ranked in order of their relatedness to certain others. For example: One kind of relationship that structures networks of documents and people is that of approval; a given judge may approve of a given document to a certain extent at a given time. Documents may thus be ranked in order of the extent to which they are approved by that judge. Of course, people may approve or disapprove of documents for many different reasons, or on the basis of many different criteria. One such criterion is the perceived informativeness of the document — i.e., the extent to which it is perceived to cause (or to be likely to cause) change of a desirable kind in the judge’s cognitive state.

However we specify the general function of a document retrieval system, the specific function of the particular component of the system that is typically known as the “retrieval mechanism” is to select an order in which to present the documents in the collection to the searcher. In other words, the function of the retrieval mechanism is to map the relatively complex network structure of the document collection to some ranked list that is simply linear rather than networked in structure. This mapping function is understood to be necessary given the assumption that the searcher, limited by the linear nature of time itself, is able to view and process a set of multiple documents only in serial fashion, one by one. The order chosen by the ranking algorithm acts as the system’s best guess as to the order that would be selected by the searcher if they were somehow to have prior knowledge of the degree to which each document is actually approved by them in the current context.

We may distinguish between two kinds of ranking on the basis of the extent to which the ranking is personalized to the individual searcher. Rankings of the first kind, Type I, are not personalized in this sense: To produce such a ranking, the system does not require the searcher to supply a query of any kind (whether pre-defined or user-defined) to which the system should respond; at a given time \( t \), the same Type I ranking would be produced whoever the searcher were and whatever
their particular needs. Conversely, rankings of Type II are personalized to the extent that their production rests on the searcher's ability and willingness to communicate a representation (pre-defined or user-defined) of their information need to the system, or on the system's ability automatically to derive such a representation from available data about the individual searcher's characteristics or prior activities.

3. Graph-Theoretic Representations of Document Networks: The CCC Model

Any document network may be considered as consisting of six sets of elements: three sets of objects — documents, term-types, and human information-seekers or "judges" — and a further three sets of object-pairs — document–term pairs, judge–document pairs, and document–document pairs. Each of the three sets of pairs may be represented in matrix form, each matrix populated with binary values. In this way, we may derive a document–term content matrix (M1), a judge–document approval matrix (M2), and a document–document adjacency matrix (M3). The binary data indicate the existence or non-existence of a pairing: in matrix M1, the presence or absence of a particular document in the set to which a particular term-type has been assigned; in M2, the approval or non-approval of a given document by a given judge; and in M3, the adjacency or non-adjacency of one document to another. If we wished to record a richer representation of the structure of the network, we could just as easily use non-binary values to indicate the weight of a term within a document description, the extent of a judge's approval, or the strength of a citation or hypertext link.

We can go on to distinguish between three basic types of retrieval system, each of which uses a different matrix as base data for further manipulation:

- content-based systems ("traditional" IR systems), whose base data is the content of a document–term content matrix (M1; Fairthorne, 1956);
- collaboration-based systems (also known as collaborative filtering or recommender systems), based on a judge–document approval matrix (M2; Resnick et al., 1994); and
- context-based systems (also known as link-analytic or hypertext IR systems) based on a document–document adjacency matrix (M3; Salton, 1963).

For convenience, we may refer to the unifying graph-theoretic model that encompasses systems of each of these three types as the CCC (content–collaboration–context) model. In all of these systems, the base data is manipulated for the same purpose (that is, in order to produce document rankings), and in the same ways (that is, using the same few, simple matrix operations).

It is possible to identify at least nine methods of producing a Type I ranking, each method utilizing a different source of base data (Fumer, in review). Two methods begin with the data given in a document–term content matrix (M1); two with the judge–document approval matrix (M2); and five with the document–document adjacency matrix (M3). The core computation in two of the methods is simple row summation; in two, it is column summation; the remaining three involve more-complex matrix operations such as the derivation of eigenvectors. These distinctions are summarized in Table 1; rankings include those by "PageRank" (Brin & Page, 1998), and by "authority" and "hubness" (Kleinberg, 1999).
Similarly, it is possible to identify eight methods of producing a Type II ranking, each method again utilizing a different source of base data. Two methods begin with the data given in a document–term content matrix (M1); two with the judge–document approval matrix (M2); and four with the document–document adjacency matrix (M3). The core computation in three of the methods is simple row comparison; in three, it is column comparison; the last two involve a more-complex matrix operation that requires the identification of the shortest paths between the nodes in a graph. These distinctions are summarized in Table 2.

<table>
<thead>
<tr>
<th>Source of base data</th>
<th>Core computational technique</th>
<th>Doc–term content matrix (M1)</th>
<th>Judge–doc approval matrix (M2)</th>
<th>Doc–doc adjacency matrix (M3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row summation</td>
<td></td>
<td>Document length</td>
<td>Judge indiscrimination</td>
<td>Document citedness</td>
</tr>
<tr>
<td>Column summation</td>
<td></td>
<td>Term frequency</td>
<td>Document popularity</td>
<td></td>
</tr>
<tr>
<td>Derivation of eigenvectors</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>Document “PageRank”, Document authority, Document hubness</td>
</tr>
</tbody>
</table>

Table 1. Typology of Type I Rankings

<table>
<thead>
<tr>
<th>Source of base data</th>
<th>Core computational technique</th>
<th>Doc–term content matrix (M1)</th>
<th>Judge–doc approval matrix (M2)</th>
<th>Doc–doc adjacency matrix (M3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row summation</td>
<td></td>
<td>Doc–doc co-indexing</td>
<td>Judge–judge consistency</td>
<td>Doc–doc coupledness</td>
</tr>
<tr>
<td>Column summation</td>
<td></td>
<td>Term–term co-occurrence</td>
<td>Doc–doc co-approval</td>
<td></td>
</tr>
<tr>
<td>Identification of shortest paths</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>Doc–doc forward distance, Doc–doc backward distance</td>
</tr>
</tbody>
</table>

Table 2. Typology of Type II Rankings

5. The UCXtra Prototype

Much about the design of search engines that provide access to networks of textual documents (collections of web pages, for instance, or scholarly papers, or library books) is well understood. Many techniques for identifying the particular documents that are relevant to the needs of a given information-seeker are known to work reasonably effectively and efficiently, and are implemented in commercial search systems that are successful and popular. Such techniques include those based variously on content, collaboration, and context. At the current time, however, it
seems that — with the prominent exception of NEC's experimental "scientific literature digital library," ResearchIndex (formerly CiteSeer; accessible on the web at <http://www.researchindex.com/>; Lawrence, Bollacker, & Giles, 1999) — no single retrieval system exists that offers the user the opportunity to take advantage of more than a few of these techniques in combination. The web search engine Google, for example, for all its noted success, employs just one method of producing a ranked list of documents in response to a searcher's query. Should the user wish to create a different ranking of documents based, say, on co-citation ("Pages that link to this one also link to") rather than on similarity of content, that option is not available.

This observation raises the following dual hypothesis: that access to networks of textual documents may be improved by allowing the individual searcher to make any decision as to the choice of retrieval algorithm that should be used in a given circumstance; and that user control over the retrieval process is most satisfying and most easily exploited when it is offered within the environment of a single, integrated system equipped with a single, homogeneous interface to multiple, diverse retrieval mechanisms, rather than with any requirement continually to switch back and forth between systems.

The central claim is not only that we may, in certain circumstances, improve the effectiveness of information retrieval by bringing multiple retrieval techniques to bear on single instances of information need; but also that we may, in certain circumstances, further improve the effectiveness of a system implementing multiple techniques by offering the user the opportunity to make her own dynamic, interactive selection from amongst the available techniques. Rather than taking either an "expert designer" or an "expert system" approach to the problem of search-technique selection, the suggestion under consideration is that we should perhaps focus on the design of systems which take direct advantage of the individual searcher's conscious ability to make reasoned decisions as to the type of search procedure that should be followed by the system in any given context or at any particular stage in the search process. This latter approach might be called the "expert user" solution, although doing so would overstate the level of expertise of the typical searcher; the "expertise" that is leveraged is simply the searcher's perception (what we might otherwise count as knowledge) of distinctions between search procedures that work well and ones that do not.

In the Department of Information Studies at the University of California, Los Angeles, we are currently working towards the development of a prototype, "proof-of-concept" system that will allow us to test the validity of these arguments. The prototype is called UCXtra, the idea being that, as the searcher, "you see extra" documents—documents that you might not get to see if you had less control over the retrieval process than this system allows. The principle guiding our design is that information-seekers may benefit not only from being given the opportunity to exploit multiple sources of evidence of document relatedness, but also from being allowed to maintain control over the ways and combinations in which such sources are exploited in any given context. However, it is becoming increasingly clear to us that the most difficult problem involved in any attempt to produce this benefit may lie in the design of the interface rather than in the technical specification of any individual method of exploitation.
6. Summary and Further Work

In this paper, I have proposed (a) a way of defining the boundaries of a unique field of inquiry (document network analysis, or DNA) whose focus is the analysis of networks of documents, (b) a unifying model of document retrieval systems (CCC) based on a specification of the document–term–searcher graph, and (c) a design for a system (UCXtra) that allows its users to take control of the process by which multiple sources of evidence of document relatedness are exploited in the course of retrieval. The next steps to be taken are the construction, implementation, and evaluation of UCXtra, thereby testing the usefulness of the approach implicit in the framework supplied by DNA and in the model supplied by CCC.

Acknowledgments

The work described in this paper is funded in part by a grant from the Academic Senate’s Council on Research, University of California, Los Angeles.

References


José Manuel Barrueco  
University of Valencia, Valencia, Spain

Vicente Julián Inglada  
Polytechnic University of Valencia, Valencia, Spain

Reference Linking in Economics: the CitEc Project

Abstract: Over the last few years, reference linking has become an active research field in digital libraries. In this paper, we present a software system that gives a solution for solving the automatic allocation reference linking. The system is based on the agent paradigm which has been applied in different approaches, in order to find a solution for the information explosion problem. The agent is called CitEc agent and it has been tested with the documents available on the RePEc data set.

1. Introduction

The relationship between documents established through citations and bibliographic references is a characteristic that differentiates scientific literature from other literary representations. Such a relationship has been investigated from multiple disciplines like scientometrics or information retrieval. The possibility of navigating the scientific literature through such relationships or to automatically access the full text of a cited document by simply clicking on its reference have been dreams for a long time. However, two requisites were needed for such dreams to become reality:

1. documents should be available in electronic format.
2. algorithms should be created to automatically extract the reference data from the document full text.

The first requisite is already achieved because more and more documents are available on the Internet. The second one is an active research field at the moment. Reference linking is the process of automatically linking each work cited in a document with its electronic full text. This is a straightforward process when metadata about references included in a paper is available as a result of the publishing process. Problems arise when it needs to be identified and extracted out of the papers full text. Several software solutions have been created to automatically deal with these problems. The agent described in this paper is one of them.

An increasing number of different applications based on the agent paradigm have appeared over the last few years in order to find a solution for the explosion of information (Julian, 2000). Taking into account other factors, Internet becomes an excellent proving ground for agent and multi-agent development. An agent is defined by its flexibility, which implies that an agent is (Wooldridge, 1995):

• Reactive: an agent must answer to its environment.
• Proactive: an agent has to be able to fulfil his own plans or objectives.
• Social: an agent has to be able to communicate with other agents by means of some kind of language.
The agent based paradigm is a suitable model to support this kind of problems because it allows an automatization of the different involved processes, an autonomy degree in order to act independently of the user, to solve complex and dynamic problems like reference linking in a flexible way and to apply intelligent techniques for locating and retrieving information.

Following this line of research, an approach to a system architecture which is based on the agent paradigm for solving the automatic allocation of reference linking data is presented in this paper. Basically, the reference linking work consists of three parts: reading, parsing and linking of documents. We have developed an agent to accomplish autonomously each one of this steps. So, we present CitEc. CitEc has been developed at the Universidad Politécnica de Valencia with the objective of reference linking documents available on the RePEc (Research Papers in Economics) data set.

RePEc (http://www.repec.org) is a distributed digital library specialised in Economics. Since 1997 is providing the latest research results in Economics. At present it holds bibliographic descriptions of more than 50,000 full text documents (articles published in journals and working papers). Most of them are freely available on the net. RePEc is based in a distributed architecture where institutions world wide share information about the documents they publish or distribute. For more information (Barrueco, 2000).

The rest of the paper is structured as follows: Section 2 shows the related work, in section 3 we present the architecture of CitEc agent. Section 4 gives the results obtained and, finally, section 5 explains some conclusions

2. Related work

Reference linking has become over the last few years an active research field in digital libraries. The publishing industry and the academic communities of electronic archives have carried out projects in this area. This section describes some of these initiatives.

OpCit (Open Citation Project) is a three years project funded jointly by the Join Information Systems Committee (JISC, UK) and the National Science Foundation (NSF, USA). One of the OpCit aims is to link the documents distributed through the arXiv preprint archive. This archive is the largest repository of scientific documents available on Internet. At the moment, with more than 150,000 documents freely available, it holds almost half of the literature produced in the field of High Energy Physics.

OpCit has developed software to automatically extract the citation information out of the documents full text and it has created links between cited and citing documents which are available on the archive. A full version of the archive with interconnected documents is running at: http://arabica.ecs.soton.ac.uk/cgi-bin/search_tj. OpCit and RePEc work on completely different environments. While in Economics there is a very large range of documents contributed by different institutions and produced with a variety of software and over different platforms, in the Physics archive almost all files are created using TeX source files submitted by the authors.

In a similar working line the CERN Document Server has reference linked their technical documents. CERN holds more than 170,000 full text documents. This initiative is similar to OpCit in the sense that both are working with
homogeneous documents. The system can be accessed at: http://weblib.cern.ch. They are offering information about the references of each document and about the cites such document has received. By contrast with OpCit, in CERN the new data has been fully integrated in the system that researchers have been using for years. As reported in (Clavivaz, 2001) near three million references have been extracted from the processed documents. For 1,937,162 of them the system has been able to create a link between the reference and the document it represents. Links have been established with documents available both in the repository and in journals subscribed by the CERN library.

CrossRef is an initiative carried out by the publishing industry. It is managed by PILA (Publishers International Linking Association) a society made up of more than 80 publishers and abstracting services around the worldwide. It is different of previous initiatives in the sense that it does not need to look into the documents to extract their references. On the contrary, publishers submit to CrossRef the metadata of the documents they publish. In this sense, is a more simple system.

CiteSeer is not just a reference linking software that could be applied to any electronic document in any discipline, it goes further in building true autonomous citation indexes. It has been developed at the NEC Research Department and described in papers like (Lawrence, 1999). A sample CiteSeer database is available at the url: http://csindex.org. It contains more than 200,000 documents with over two millions references. CiteSeer is a free of charge distributed software for non-commercial institutions. It can be used to create repositories of scientific documents and citation indexes following the model of the ISI Citation Indexes. It is able to locate scientific documents on the Web, to download the documents full text in order to extract their list of references, to split each reference in different elements (title, author, etc). It can also determine if two references with different format represent the same document, etc.

CiteSeer is relevant for our purposes because it works in environments similar to RePEc, since it deals with heterogeneous sets of documents that have been created with a variety of software, with different formats and more important of all, with different citation styles. Each discipline has a particular citation style, which is employed by almost all authors. In Physics, for instance, a typical reference is quite simple because the work title is omitted and the source data is abbreviated. It could look like:


Nevertheless in Economics references use to be more complex. Most of them use the Chicago Manual of Style (Chicago, 1993). That makes software developed by CERN or OpCit unusable in our dataset.

3. The CitEc agent

CitEc is an agent designed to autonomously operate on the RePEc universe of data, but it can be applied to other information sources too. It basically has three functions: it reads documents in RePEc, it extracts their citation data and, finally, it links references with the full text of the documents they represent if are available on RePEc. CitEc is structured as different modules with different tasks. Its architecture is presented in the following point.
3.1 Architecture

It can be seen from the chart in figure 1 that CitEc is made up of three components: the processing subsystem, the knowledge base (KB) and the communications interface.

The most important part of CitEc is the processing subsystem. Basically it is in charge of detecting new full text documents, downloading and parsing them in order to find out their lists of references and finally linking the references with the full text of the documents they represent. Each task is executed by one of the following modules: reading, parsing and linking.

The CitEc inputs are templates containing bibliographic information about scientific documents. The CitEc agent processes this data and, as a result, it modifies the RePEc environment by adding a new template type called "citation template" (Barrueco, 2002). There will be a citation template for each document that has been successfully processed. It will include the list of references with the different identified elements of each one. In addition, a link from the cited document will be included if it is available on RePEc data set.

The parsing module is the core part of the agent. It accomplishes three main tasks. Since RePEc holds only bibliographic data with pointers to the full text of the documents, the first operation performed is to download the files containing the full text of the documents. Files in other formats than PDF or PostScript are excluded. This is a little constraint because more than 94% of documents in RePEc are made available in one of these formats. Secondly, it converts the files from the original format to ASCII, so the text can be parsed. CitEc uses \texttt{pstotext} for this operation. By far, this is the most complicated task since there is no a reliable application to convert PDF files to ASCII. See (Robinson, 2001) for a comparison of conversion tools. Finally, the last task of this module is to parse the ASCII file in order to find out if the author has included a list of references. We have used in this task CiteSeer procedures rather than to design new ones. CiteSeer has been probed as a very
successful tool in the citation analysis of heterogeneous sets of documents (Lawrence, 1999). Once the references have been extracted, the module tries to isolate each individual reference and then to parse it in order to find out its elements. The module is able to identify the publication year, authors, title and context in which the citation appears in the body of the document.

Finally, the linking module reads the information about references extracted by the previous module. For each reference found, it tries to identify if the document is available in RePEc. It compares the reference elements against the KB. If it finds a similar document, a link is established between both documents. This module is also in charged of modify the RePEc environment by adding new templates to the data set. Each time a reference is read and processed a new template is created.

The second component of CitEc is the knowledge base. It maintains an internal representation of the RePEc data set. As described in (Barrueco, 2000), RePEc is based on a relational structure that can be easily translated to a SQL database. In this way the KB of CitEc has been developed as a relational database and has been implemented using a MySQL database management system. The KB contains three basic tables: DOCUMENT is a table that holds the basic bibliographic information for each document available in RePEc. REFERENCE is a table that contains a row for each reference extracted of the documents. Finally, CITATION is a table that contains a row for each relation established between cited and citing documents in RePEc. It has two columns, the first one holds the key of the citing document and the second one the key of the cited document.

Finally, the communications interface provides an easy way to interrogate directly the KB from third part applications. This module accepts the methods described by (Bergmark, 2000) in the Cornell reference linking API. It is implemented as a CGI script available at: http://netec.ier.hit-u.ac.jp/adnetec-cgi-bin/get_data.pl. This script requires two arguments. The first argument is a handle of the document which we need to get information of, and the second argument is one action. There are three implemented actions: getMyData returns the document metadata, getReferenceList gets the list of references of the document and getCurrentCitationList returns the list of documents which have cited the current document.

4. Results

CitEc has been on since January 2002. Over this period of time, 28157 documents have been analyzed. 13446 (48%) have been successfully processed. 293069 correct references were extracted out of them. This figure gives us an average of 22.5 references by document. 52352 (18%) references represented a document available in RePEc, therefore a link between both documents was created. Nevertheless, for the purposes of this study is more important to look into the errors.

It can be seen from the chart in figure 2 that the most important cause of error is conversion from PDF/PS to ASCII. Since pstotext has a high rate of errors, more sophisticated conversion tools are required. Nevertheless this point is out of our scope. Secondly, CitEc has been unable to find a list of references in the 12% of documents. This figure is quite high because it is not usual to find scientific documents without a bibliography. We have detected that in some cases the
problem behind this error is that *psiotex* crashes before to finish the process. That produces a correct but uncompleted ASCII file. Since references are usually placed at the end of the paper, CitEc can not find a list of references.

Other problems with lesser incidence are documents with wrong number of references. We consider that the process to isolate references is correct when it finds between 1 and 75 references. If the number is out of these limits the document is discarded. Another cause of errors is documents which have been distributed in other formats than PDF or PS. As we can see this is a minimal part of documents. Moreover, there is a 3% of documents, which are not in English language. This constraint will be removed in future versions. Finally, since the public release of CitEc on January, 14th until March, 9th, it has got 4674 queries (77 request by day). At the moment, the results of our work have been implemented in three RePEc services: WoPEc: (http://netec.mcc.ac.uk/WoPEc), SocioNet (http://socionet.ru) and BibSoc (http://www.uv.es/bibsoc/GM).

5. Conclusions

An approach to a system architecture which is based on the agent paradigm has been presented in this paper. This approach gives a solution for solving the automatic allocation reference linking. CitEc is an agent that automates the process of reference linking between different documents in RePEc. It allows to automatically obtain information about the references of a document in a digital library. CitEc has been tested with a great quantity of documents and it seems to work well to reduce enormously the time needed in this process.

As future works we want to reduce the error percentage of the evaluated documents in the different modules of the CitEc agent. Another direction for future work is to improve the research algorithms of the references.
References
Ailyson Carlyle and Lisa M. Fusco
University of Washington, Information School, Seattle, WA, USA

Equivalence in Tillett’s Bibliographic Relationships Taxonomy: A Revision

Abstract: This paper analyzes the equivalence relationship as presented by Barbara B. Tillett in her taxonomy of bibliographic relationships. Tillett’s definition of equivalence comprised of two parts, first, that equivalence holds between exacts copies of bibliographic items or documents, and second, that it may hold between an original item and a reproduction, if the intellectual content and authorship are preserved. It is proposed that this definition is too restrictive, excluding relationships among items that may, based on contexts of use, act as equivalent. Further, it is suggested that a taxonomy of bibliographic relationships be constructed as holding between document representations as opposed to documents themselves. A revised definition of equivalence is offered in which equivalence relationships may hold among document representations in which one or more document properties described in the representations are shared. One advantage of this revision is that it subsumes Tillett’s shared characteristic relationship, simplifying the taxonomy.

1. Introduction
Barbara B. Tillett’s taxonomy of bibliographic relationships uniquely addresses a long standing concern of the cataloging community, namely the identification and use of relationships among items identified in library catalogs. Tillett states that one of the major purposes motivating her development of the taxonomy is its subsequent capability to “act as a rationale to guide in the creation and use of the catalog” (Tillett, 1987, p. 23). We suggest that, given the increase in document types and contexts of use, the purpose be extended beyond library catalogs to include a wide variety systems of knowledge organization. We further propose a shift in emphasis moving away from the identification of fixed or stable relationships that exist among documents independently of their use and toward a view of relationships that recognizes their variability given the changing nature of uses, users, and contexts. In light of this extension of purpose and emphasis, we differ with Tillett on the strategic roles played by relationships types in the design and use of such systems. We suggest a revision of Tillett’s taxonomy to accommodate an expansion of purpose and emphasis that will meet the needs of today’s users, including users of the library catalog. This paper will concentrate on the redefinition of one of Tillett’s relationship types in order to exemplify the rationale and strategy for revision. The relationship type offered as the basis for discussion is that of equivalence.

2. Tillett’s Bibliographic Relationships Taxonomy
In the introduction to her dissertation, Bibliographic relationships: Toward a conceptual structure of bibliographic information used in cataloging (1987), Tillett states her intention as follows:
to provide the groundwork for understanding the conceptual structure of the ideal library catalog in terms of bibliographic relationships (p. xx).

She then provides a definition of conceptual structure as a framework, comprised of the following three parts:

1. what should be included in the catalog;
2. the elements necessary to describe these items; and finally
3. the links or relationships among the items (p. 6).

Tillett’s research addresses only the last of these three parts. Specifically she intends to “operationally [identify] and [categorize] bibliographic relationships in a taxonomy” (p. xx). The bibliographic relationships identified and categorized in the taxonomy are:

1. equivalence relationships, which hold between exact copies of the same manifestation of a work or between an original item and its reproductions, as long as the intellectual and artistic content and authorship are preserved...
2. derivative relationships … which hold between a bibliographic item and a modification based on that item …
3. descriptive relationships, which hold between a bibliographic item or work and a description, criticism, evaluation, or review of that work …
4. whole-part (or part-whole) relationships … which hold between a component part of a bibliographic item or work and its whole …
5. accompanying relationships, which hold between a bibliographic item and the bibliographic item it accompanies…
6. sequential relationships … which hold between bibliographic items that continue or precede one another ..
7. shared characteristic relationships, which hold between a bibliographic item and other bibliographic items that are not otherwise related but coincidentally have a common author, title, or other characteristic used as an access point in a catalog … (Tillett, 1991, p. 156).

Although Tillett develops her taxonomy through an analytical and empirical study of cataloging records and codes, the definition of the conceptual structure focuses on the items described in cataloging records and linked by the relationships identified in those records rather than the representations or records of the items thus described and linked. Tillett herself, in a discussion of her methodology for identifying types of bibliographic relationships, states that her taxonomic categories are derived from both the definitions of types of physical items found in glossaries and inferences made from her reading of cataloging rules. In other words, although Tillett’s taxonomy ultimately categorizes bibliographic relationships based upon an analysis of their representation in cataloging records, the categories themselves are derived from definitions of types of physical items gleaned from cataloging codes and glossaries.

In Tillett’s view, the relationship types resulting from the grouping of items can subsequently function as pathways linking related items based on the bibliographic characteristics incorporated into cataloging records and shared by items in those categories. In addition to functioning as road signs, the links themselves display to the user the availability of related items. The taxonomy of bibliographic relationships thus aids in permanently stabilizing relationship types with reference to those items solely representative of each type.

The purpose of our proposed revision Tillett’s equivalence relationship is to identify sets of properties or attributes brought forward in document representations
that exemplify relationships and can be re-grouped and re-used to create equivalent groupings of document representations dependent upon varying material contexts of the use of the documents themselves. Multiple uses of documents result in the need for developing criteria of their substitutability. Thus, rather than using the equivalence relationship to conventionally link identical items, providing users pathways among related documents, we redefine equivalence to accommodate the notion of substitutable representations of the same document altered by the context of its use. Our proposed revision of Tillett’s equivalence relationship is designed to address the unfixed and impermanent identity of a document as articulated by its use.

Central to our proposed revision of Tillett’s taxonomy of relationship types is the idea of substitutability. Basically, substitutability is a condition of document usage; the properties shared by these documents are relative to that use. For example, there are times when one edition is as good as another; times when we need to be alerted to various editions to suit our purposes; and, times when only one particular edition will meet our needs. It is our contention that any optimally functional knowledge organization system, including the library catalog, must be capable of accommodating users’ specifications of those properties identified as significant in their determination of substitutability. The redefinition of the equivalence relationship is a step toward the creation of responsive, user-centric, and context-sensitive knowledge organization systems.

3. Tillett’s Equivalence Relationship

As noted above, Tillett defines the equivalence relationship as follows: “Equivalence relationships are those that hold 1) between exact copies of the same manifestation of a work, or 2) between an original work and reproductions of it, as long as intellectual content and authorship are preserved” (Tillett, 1987, p. 27). No explicit criteria are stated for determining when the proviso holds between an original work and its reproduction. The implication is that the decision must be made repeatedly in any given number of circumstances. In other words, the ground for determining this particular type of equivalence is situational. Equivalence relationships, then, according to this second proviso, hold in various contexts of decision-making, rather than between documents that are potentially duplicate with regard to the preservation of intellectual and artistic content. In 2001, Tillett acknowledges that “the idea of equivalence is tricky, because it is a subjective determination whether entities are indeed equivalent or not, depending on what the user considers to be the ‘same’” (Tillett, 2001, 24). However, she does not, based on this acknowledgement, revise her definition of equivalence.

4. Equivalence Relationship Revised

Our proposed revision of Tillett’s taxonomy of relationship types suggests that equivalence be determined syntagmatically; that is, that it be defined relative to the use of documents. In addition, we propose that the distinction between item, or document, and representation of that item or document, be clarified and articulated. Not all attributes of a document are equally significant in every context of use. Further, not all are identified in a representation of that document. Given this, we propose a model of equivalence relationships that allows for the selective
distribution of attributes based on (a) contextual states of substitutability between documents and (b) the presence of those attributes in a document representation.

One of our motivations for redefining the equivalence relationship stems from a consideration of the collocating objective of the library catalog as conceived by Seymour Lubetzky and of the current activity surrounding the publication of a conceptual model of bibliographic entities in IFLA's *Functional Requirements for Bibliographic Records* (IFLA, 1998). Lubetzky stated that the collocating objective of the library catalog required the catalog "to relate and display together the editions which a library has of a given work and the works which it has of a given author" (Lubetzky, 1960, ix). We suggest that one of the means of "relating and displaying together" editions of a work and works by an author in a library catalog or information system is to treat them as equivalence sets and group representations describing those editions and works based on the identification of shared properties that exist in the cataloging records.

In *Functional Requirements for Bibliographic Records* (*FRBR*), an entity-relationship model of the bibliographic universe is proposed to "provide a ... shared understanding of what it is that the bibliographic record aims to provide information about" (IFLA, 1998, p. 2). *FRBR* posits four bibliographic entities:

1. *work* ("a distinct intellectual or artistic creation"),
2. *expression* ("the intellectual or artistic realization of a work"),
3. *manifestation* (the physical embodiment of an expression of a work"), and
4. *item* ("a single exemplar of a manifestation").

In the *FRBR* model, each of these entities is associated with particular attributes that would be described in cataloging records. For example, a cataloging record for *The Haunted Pool*, Frank Hunter Potter’s translation of George Sand’s *La Mare au Diable* published by Shameless Hussy Press in 1976 would include attributes for the work, expression, and manifestation entities, as follows:

1. *work*: George Sand (author name attribute); *La Mare au Diable* (work title attribute)
2. *expression*: Frank Hunter Potter (translator name attribute); *The Haunted Pool* (expression title attribute)
3. *manifestation*: Shameless Hussy Press (publisher name attribute); 1976 (publication date attribute).

Because the model is hierarchical, work attributes are inherited by expressions; work and expression attributes are inherited by manifestations; and work, expression, and manifestation attributes are inherited by items. We suggest that if equivalence is defined to hold among document representations sharing a selection of properties or attributes, then the records associated with particular works or expressions could be regarded as sharing an equivalence relationship based on the properties they have in common. Thus, all of the records that share the author name "George Sand" and the work title "La Mare au Diable" could be considered equivalent in a particular context of use, and could be displayed together as a unit or set in an information system. Using the *FRBR* model, a group of records sharing expression attributes could also be displayed together to identify the members of an expression set.

This view of equivalence varies from the Tillett view in that the membership of an equivalence set may change depending on variant contexts of use. Those contexts of use may be identified by the cataloging community, as is the case with the *FRBR* model, or they may be identified by individual users. Thus, an
expanded view of the equivalence relationship based on shared properties or attributes is amenable to alternative uses of document representations.

A redefinition of equivalence based on shared properties would not only accommodate predominant viewpoints such as those articulated by Lubetzky and FRBR, but would accommodate viewpoints articulated less frequently that may hold in non-typical user contexts. For instance, in some contexts, for some users, an item that is related to another work may be considered to be equivalent to the work itself. One example of this may be a context in which a movie version of a textual work would be considered to be equivalent to, or substitutable for, the textual work itself.

Based on a view of equivalence as a syntagmatic relationship to be called out by the specification of shared properties identified in document representations, we propose the following definition of equivalence relationships shown in knowledge organization systems:

Equivalence relationships hold among documents representations in which one or more document properties described in the representations are shared.

This redefinition of equivalence relationships makes clear the fact that equivalence relationships hold and are made possible in document representations. The shift from documents to document representations is a useful one given the nature of documents and the nature of document representation. It sometimes happens that documents do not state the properties that they embody that would identify them as equivalent to other documents. For example, a document that is a translation of another document may not state the title of the document of which it is a translation. In this case, a critical property that could be used in establishing equivalence is missing. In the library cataloging environment, that property is supplied in the cataloging record if it is obtainable. Thus, document representations sometimes have more power in establishing equivalence relationships than the documents themselves.

Another advantage of the redefinition proposed here is that equivalence relationships may be used to describe entities identified in models of the bibliographic universe proposed elsewhere. For example, the bibliographic entities described in the FRBR model could be easily accommodated, and works and expressions, not as easily accommodated in Tillett's existing taxonomy, could all be identified using the equivalence relationship.

Redefining equivalence based on shared properties allows us to further simplify Tillett's relationships taxonomy. In the taxonomy proposed by Tillett, one of the relationship types is shared characteristic. Given a redefinition of equivalence based on shared properties, shared characteristic relationship is no longer necessary, and is subsumed in the equivalence relationship. One advantage of this is that relationships among two different works that do not share a derivative relationship but share a common origin, such as the relationship between the movie Scrooged, based on Charles Dickens' A Christmas Carol, and the children's picture book produced by Disney, Mickey's Christmas Carol, are accommodated by the same relationship that would show the connection between these two works and the original by Dickens.

5. Conclusion

In contrast to the intention of Tillett's taxonomy of bibliographic relationships to provide for a greater understanding of the conceptual structure of
the catalog, this paper offers, as an alternative, a conception of a taxonomy as a way to strengthen the *theoretical foundation* of a knowledge organization system as articulated in and implemented by its objectives. To do this requires a redefinition of the equivalence relationship identified by Tillett’s taxonomy of bibliographic relationships. Specifically, the paper demonstrates the ways in which the proposed redefinition serves to support users in their searches for bibliographic items represented in information systems.

**Acknowledgements**

This paper represents a slice of current thinking that is part of an ongoing exploration of Tillett’s taxonomy of bibliographic relationships being conducted by the authors and David M. Levy. It would not have taken place without the pioneering work of Barbara B. Tillett, to whom we and the knowledge organization community in general are much indebted.

**References:**


Abstract: A revision of the taxonomies of the bibliographic relationships is made settled down by Tillett, Leazer or the study Functional Requirements for Bibliographic Records of the IFLA. A group of bibliographic records of related works is analyzed, that we have denominated bibliographic families, in the union catalogue of REBIUN (Net of University Libraries). The techniques more used by the cataloguing rules and the IBERMARC Format are identified to show the derivative relationships in the catalogue and it is concludes that many of those "kinship relationships" existent in the bibliographic families they are not made explicit for data elements and the access points of the bibliographic records.

1. The bibliographic relationships in the catalogue.

1.1. Bibliographic entities, works and items.

One of the problems of some studies of use of the catalogue of the library that try to determine the group of elements of necessary data in the bibliographic records is, according to Holley, that consider the habitants of the bibliographic universe as discretes entities, more than as members of a galactic interconnected system. In consequence, they don't consider the data elements whose main function is to organize the catalogue, that is to say, to provide structure (Holley, 1992, p. 65-66).

Our current vision of the conceptual structure of the catalogue can be traced at least already from a date as distant as mid-nineteenth century when Panizzi defended strongly the library catalogue as the vital instrument to access a library's holdings. Their rules reflect a conceptual structure based on the entries, that is, single and full records for each catalogued item, and three classes of cross references to link the user's choice with the catalogue entry. The conceptual structure of the catalogue consists, therefore, "records", composed of "data elements" and of "links" among the records. Linked records form "clusters" that share a particular type of relationship. Records are elaborated to describe bibliographic entities and to control subjects and the names used as access points, including personal, corporate, conference and geographical names, as well as the titles, mainly the titles of works or component-parts of works (Tillett, 1989, p. 150).

A renovated attention toward the nature and the characteristics of the informative catalogue unit has come determined by the potentiality of the new technologies (and for the logic of the computer) of decomposing the monodimensional rigid structure of the catalogue data and of articulating a complex net of relationships around an informative basic unit, with the purpose of
determining a descriptive format constituted by data integrables according to the bibliographic attributes that want to be indicated in having response to the users' certain demands (Grimaldi, 1993, p. 134).

Nevertheless, in spite of the fact that more than two decades from have already lapsed the incorporation of the computers in the library tasks, the transition of the catalogue of cards to the on-line catalogues is still in its initial stages. As Svenonius points out (2000, p. 63-64), this slow transition is due in part to difficulty in understanding the function of the bibliographic record in an electronic environment. Bibliographic records in MARC format are fairly well suited to communicating bibliographic data but they are not well designed for computer manipulation of this data.

A bibliographic record is a description of a bibliographic entity that is formatted and inscribed on a medium. A bibliographic entity is a unique instance of recorded knowledge, e.g., an opera, a play, etc. Each bibliographic entity has both physical and intellectual properties. The item is the physical property represented in the catalogue by a transcribed record of its inherent bibliographic characteristics (the item's dimensions, its details of publication, etc.). The work is the intellectual content of a bibliographic entity (Smiraglia and Leazer, 1999). Any work has two properties: a) the expressed propositions; and b) the expression of those propositions (usually a certain group of linguistic or musical strings, etc.) which forms linguistic content. Any change in ideational or linguistic content results in the creation of a new work.

Descriptive cataloguing generally intends to provide the user with information about: a) the item described as physical entity, and b) work(s) contained in the document, i.e., the item's intellectual content. To differentiate between these two aspects, it the term "document" is usually applied to the item described as physical entity and the term "work" to the intellectual content of the item.

1.2. Types of bibliographic relationships.

A bibliographic relationship is an association between two or more works or bibliographic documents. Some examples are the publication circumstances that link two or more bibliographic documents, as the attribution to the same author, to contain a variation of the same work, or documents that are part of the same series.

For the automated catalogue to bring together related items, the works within a collection must be identified systematically within the cataloguing database. When a computer retrieves items according to a specified search strategy, it separates the collection into two distinct groups: those that satisfy the search and those that do not. If the search is very specific and only a few items are retrieved, it may not be necessary to order these items further. One may only need to scan the bibliographic data to choose the desired items. The retrieved items, however, are not necessarily unrelated to the items not retrieved. A more general search strategy will retrieve some of these items, but as the strategy is generalized, the relationship between any two items can become weaker. Retrieved items should be displayed to represent the varying interrelationships among them (Van Houten, 1981, p. 370).

For their own nature, some relationships are more tenuous and more subjective than others. For example, it is much easier to judge if a work is a translation than if it is a parody (Holley, 1992, p. 61). Also, the bibliographic relationships can be multidimensional. For example, the second edition of a work can be published in a different series different from the first edition, a reprint of a
sound record of rock can contain a new sound replacing the sound track of the original record and a collection of microforms often compiles material of several publications (Hagler, 1997).

Probably the most significant work on the bibliographic relationships is the definition of relationships in the UNIMARC Format (1977 and 1980). The definitions of bibliographic relationships found in the UNIMARC Format (1980, p. 58-59) suggest a philosophical framework for the bibliographic relationships by categorizing them into the following three types:

1. **Vertical**: they are the hierarchical relationships of the whole to their parts and the parts to a whole, for example, downward links (a serial to its subseries or to individual volumes of the series) or upward links (the individual volume to its subserie and/or series).

2. **Horizontal**: they are the relationships between versions of a document in different languages, formats, media, etc.

3. **Chronological**: they are the relationships in the time between issues of a document, for example, the relationship of a serial to its predecessors and successors.

Barbara Tillett carried out an empiric study during the years 1984-1986 that it constituted her doctoral dissertation. The author has developed a taxonomy that "can be seen as an expansion and clarification of the categorizations of the original UNIMARC" (Tillett, 1988, p. 5). The seven types of bibliographic relationships derived from this study are the following ones:

1. **Equivalence relationships**: they are given between exact copies of the same manifestation of a work, or between an original document and their reproductions, as long as the intellectual and artist content and authorship are preserved.

2. **Derivative relationships**: they are given between a bibliographic document and a modification based on the same document.

3. **Descriptive relationships**: they are given between a work or bibliographic document and a description, criticism, evaluation or review of that work, such as between a document and a review describing it. Also included are annotated editions, commentaries, critiques, etc. (Tillett, 1991b, p. 156).

4. **Whole-part (or part-whole) relationships**: they are given between a component part of a work or bibliographic document and its whole, as with an individual selection from and the whole anthology, collection or series (Tillett, 1991b, p. 156).

5. **Accompanying relationships**: they are given between a bibliographic document and the bibliographic document it accompanies as, for example, between two documents where one is predominant and the other subordinate or where documents are of equal status.

6. **Sequential relationships**: they are given between bibliographic documents that continue or precede one another, as between the successive titles of a serial, the sequels of a monograph or among the various parts of a numbered series (Tillett, 1991b, p. 156).

7. **Shared characteristics relationships**: they are given between a bibliographic document and other bibliographic documents that is not otherwise related but coincidentally has a common author, title, subject or other characteristic used as access point in a catalogue, such as a language, publication date or publication country (Tillett, 1991b, p. 156). This type of relationships is the most
general of all, because they are given whenever an access point is duplicated in a certain file (Tillett, 1991a, p. 402).

Barbara Tillett’s enumeration almost covers the whole spectrum of bibliographic relationships that can be given at item level. Vellucci (1997) examined six of these seven categories in their study of music’s catalogs. Zagorskaya (2000, p. 16) intends to add other two types of relationships:

8. Relationships concerning the types of material or genre (common title or form).

9. Relationships between documents with a common historical background (for early printed books: collections, ex-libris, bindings, marginalia, etc.).

The model offered by the IFLA Study Group on Functional Requirements for Bibliographic Records in 1996 (FRBR, 1998) also describes bibliographic relationships on the bibliographic record component level, as it is based both on general catalogue functions (search, identification, selection and retrieval) as on bibliographic record functions. The model contains three “entity” groups and its three corresponding bibliographic relationships groups. These are:

(a) entity representing the products of scholarly or artistic creative effort (works, expressions, manifestations, and items/documents),

(b) entity intellectually responsible for the products (persons and corporate bodies),

(c) entity forming the subject of creative effort (concepts, objects, events and places).

Bibliographic relationships between these entities are represented on the following three levels:

1. Relationships between elements within a bibliographic record;
2. Relationships between bibliographic records, implemented via access points;
3. Relationships between bibliographic records, implemented by catalogue complexes.

1.3. Super records, superworks and bibliographic families.

As an alternative to the traditional bibliographic records based on the item as basic unit of description, Fattahi (1996) proposed two kinds of "super records", constructed for the effective discharge of the collocating function: one for voluminous authors containing the author heading and titles of the works by that author; and one for voluminous titles which have different editions and manifestations.

Super records for works are a device for incorporating the concept of superwork, i.e., the set of all documents deriving from an ur-work (Svenonius, 2000, p. 98). Smiraglia and Leazer (1999) use the expression bibliographic family to designate a set of related bibliographic works derived from to common progenitor. From the perspective of the FRBR, the work is the beginning of the bibliographic family, the composition of one or more texts that are "head of the household" of the successive members of the family (Grimaldi, 2000).

2. Visibility of the kinship relationships: a concrete study.

The works or items that are part of the bibliographic families are usually represented in the catalogue by independent records. The derivative relationships or “kinship relationship” among these records are shown to the users that consult the
OPAC by means of the data elements of records and the form in that are presented. This, in turn, comes determined by the standards and cataloguing formats used in the libraries.

2.1. Objectives of the study.

This study is to determine until point the current standards "show" sufficiently to the users the "kinship relationships" existent among the related works (works published simultaneously in two editions, works that are revised with posteriority to its first edition, different translations, adaptions, representations of the same work, etc.).

2.2. Methodology.

We have carried out an intentional sampling (that is to say, the bibliographic records of works have been selected that maintain to each other kinships relationships of interesting for our study) in the union catalogue of REBIUN and an exhaustive analysis of all the elements, fields and subfields susceptibles of "to guide" to the user of the catalogue of one to another related record.

2.3. Results.

The preliminary results indicate that the techniques more used to represent the derivative relationships are the points of access of the related records (uniform titles, subjects headings, common main headings and added headings: fields 1xx, 6xx and 7xx of the IBERMARC Format), edition mentions (subfield $a$ of the field 250 of IBERMARC Format), references of the authority system, etc. Nevertheless, the complexity of many of the "kinship relationships" existent in the bibliographic families (and that can be implicit in the conceptual maps of the users that carry out a search in the catalogue) are not made explicit for the data elements of and the points of access of the bibliographic records. Furthermore, the different practices of cataloguing of the university libraries contribute in certain cases to hide these kinship relationships.

3. Conclusions.

In conclusion, we consider it necessary that those responsible for the departments of technical services of the Spanish libraries systems introduce some modifications in the cataloguing policy to give more relief to the attributes that are good to show these existent relationships in the bibliographic families. In the longer term, it is considered indispensable the realization of studies of the behavior of the users that contribute more information on the "sailing maps" that use in their searches in the catalogue, later on, carry out a revision of the cataloguing rules and of the bibliographic and authorities formats that facilitate quicker and more effective identification and retrieval of works with complex bibliographic relationships. In this sense, we think that the study *Functional Requirements for Bibliographic Records* is a good starting point for this revision.

Notes

1. Part of this work has been elaborated in the mark of the investigation project *la estructura conceptual del registro bibliográfico: fundamentación teórica y estudio de un caso* (SA082/01 indexes), subsidized by the Consejería de Educación y Cultura de la Junta de Castilla y León and the Fondo Social Europeo (Orden de 6 Febrero de 2001, B.O.C.yL. de 15 de Febrero de 2001).
2. REBIUN (Net of University Libraries). The on-line catalogue REBIUN is the last project developed by the Spanish universities with the objective that the combined funds of all them are unified and available the 24 hours of the day. Their upgrade is carried out every two months, and at the present time it contains the holdings of 52 Spanish universities. It can be consulted in the following address: http://www.crue.org/cgi-bin/rebiun.

3. An illustrative example is the cataloguing of the edition in microform of the doctoral dissertation *Análisis del comentario de Proclo a Crátito de Platón* of Joaquín Ritoré Ponce. In the catalogue of REBIUN two bibliographic records of the same item appear. Only in one of the universities a added entry of author-title has been taken out *Crátito* of Platón.

References:


UNIMARC: *universal MARC format* (1980). Recommended by the IFLA Working Group on Content Designators; set up by the IFLA Section on Cataloguing and the IFLA Section on Mechanization. 2nd ed. rev. London: IFLA International Office for UBC, 1980.


A Comparative Study of Six French-language Web Directories

Abstract: This paper presents a comparative study of six French-language Web directories (MSN, Nomade, Open Directory, Voila, Voila Pages Perso and Yahoo). The study focuses on the quantitative and qualitative aspects of the organization of these directories, and on the way in which they describe sites. It reveals a wide variety of structures, content and organizational principles. In this respect, Web directories do not correspond to classic theories of classification. They highlight the difficulty of proposing a structured representation of the heterogeneous content of the Web.

1. Introduction

Web directories, lists of web sites classified into categories, are extensively used by Internauts. They are particularly useful for novices since they facilitate navigation and provide particularly relevant links, especially compared to "full text" search engines with their problems of complexity of queries, noise, etc. The creation of Web directory categories and the classification of sites are done manually, which contributes to an acceptable organization of information: categories are created "rationally" and systematically, and the relevance of listed sites is generally guaranteed.

We undertook a comparative study of six Web directories partially or totally dedicated to the French-language Web: MSN, Nomade, Open Directory, Voila, Voila Pages Perso and Yahoo. This study concerned data available in these Web directories in February 2001. We first developed a specific software package in order to explore the structure and content of directories (hierarchical links and cross-references between categories, listed addresses and site descriptions). On the basis of these data we then performed a qualitative analysis of the organizational principles of each directory, followed by quantitative investigations consisting of: 1) calculation of statistical indicators representing the structure and complexity of each directory; 2) calculation of the specific characteristics of each directory, based on the content and presentation of the indexed sites. Our study included other analyses based on a formal exploitation of directories as graphs, but the limited scope of this article does not enable us to include those aspects as well.

2. What is a Web directory?

A Web directory offers users lists of sites grouped into hierarchically structured categories forming a "tree". These categories contain the indexed sites or pages (identified by their URL, i.e. their Internet address), along with a brief description of their content.
Directories differ not only in size and number of URLs presented and indexed, but also in structure. The structure of a directory can be defined by the combination of three elements:

Multi-indexing: some directories index the same URL in different categories; thus, the same address sometimes appears several times in different places in a directory.

Position of the URLs indexed in the tree: some directories propose URLs in all their categories, while others classify them only in terminal categories (i.e. those which do not have sub-categories).

Use of cross-references: directories such as Yahoo propose links, within categories, to other categories which may be situated anywhere in the directory and not necessarily directly beneath them in the tree.

Structurally, each directory is thus a combination of these three elements, as presented in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>Multi-indexing</th>
<th>URLs are indexed only in terminal categories</th>
<th>Use of cross-references</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSN</td>
<td>✓</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Nomade</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>Open Directory</td>
<td>x</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>Voila</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Voila PP</td>
<td>✓</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Yahoo</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 1 - Structural description of directories

3. Differences of size and structure

The first difference we observe is in the coverage of directories, calculated in terms of the number of indexed URLs. This calculation must take into account multi-indexing: if a directory lists the same URL in several places, it will present more addresses than individual URLs actually indexed. That is why it is important to distinguish between the number of URLs presented and the number of individual URLs indexed (Table 2). For example, whereas Nomade is the directory with the broadest coverage (139,000 indexed URLs), Yahoo presents the highest number of URLs to the user and thus has the highest repetition rate.

Directories also vary highly in terms of depth. Yahoo, with its 18-level depth tree, has the most depth. The smallest directory, Voila Pages Perso, has a maximum depth of only 5. But depth does not systematically correlate with the number of URLs (see Table 2). We furthermore found wide variability in terms of density (mean number of URLs per category). Whereas Nomade and Voila propose a mean of close to 20 URLs per category containing at least one URL, Open Directory, Yahoo and MSN offer a mean of between 5 and 10.
Finally, cross-references strongly influence the directory structure. They facilitate navigation for the user and enable the creators of directories to compensate for the rigidity of the overall organization of the tree. The four directories that use cross-references (Nomade, Open Directory, Voila and Yahoo) do so differently (see Table 3). While Nomade and Voila make little use of them (only 1.6% of Voila categories use cross-references, with a mean of 1.4 cross-references proposed), Open Directory and Yahoo use them on a big scale (20% of the Yahoo categories, with a mean of 4 cross-references per category).

4. Web directories have different organizational principles
In this section we consider the principles governing the organization and structuring of directories. Several models can be used for the organization of data
and knowledge, derived from fields as varied as knowledge representation in Artificial Intelligence, the compilation of thesauruses (e.g. in libraries) or the creation of indexes and other "paper" directories (e.g. yellow pages, professional directories). These different models have been adopted by the publishers of Web directories, in some cases intentionally, in others less so. Three modes of organization can be distinguished:

- Systematic categorization of domains of human activities, objects of daily life, etc. in an ontological approach.
- Less systematic and more practical cataloguing, focused on human activities (e.g. business, recreation, various forms of sociability), in a "yellow pages" approach.
- Categorization of the "Internet world": mapping of sites and services available on the Internet, without any precise criteria for classification of objects of the world, human activities, etc.

Note that Internet directories, at least those that we studied, do not correspond strictly to any one of these approaches; they combine the principles of usual classificatory objects: ontologies, thesaurus, etc. In the six directories that we studied, very different organizational models were identified.

For example, if we compare Yahoo and Voila, Yahoo has a systematic classification approach evident in a large number of categories (44,000 compared to 9,000 in Voila) organized in an 18-level tree structure (compared to 10 levels in Voila). Yahoo also has a very dense network based on a system of cross-references between categories (34,000 cross-references, compared to 200 in Voila). The main first-level categories in Yahoo are "Regional" (Exploration géographique) and "Business and economy" (Commerce et économie). This reveals a systematic classification approach. The encyclopaedic aspect of the Yahoo directory is also evident in the presence of categories such as "Social science" (Sciences humaines) at the top of the tree. By contrast, Voila has a pragmatic approach, focused on services related to human activities: business, social activities and recreation. This practical aspect of Voila is strengthened by the existence of a first-level category "Shopping, daily life" (Achat, vie pratique) which accounts for 13% of all indexed sites and has no equivalent at the top level in Yahoo France directory.

5. Web directories have little in common

The six directories studied index a total of 283,000 distinct URLs. We found that in general the directories overlap very little and have only 25 URLs in common; 70% of all the indexed URLs are in only one directory and 89% in one or two directories.

Each directory has its specific characteristics, confirmed by a low two-to-two overlap. Since the sizes of directories differ, the calculation of two-to-two overlaps between directories is asymmetrical and must be analysed for each pair of directories (Table 4).

The particularity of the Voila Pages Perso personal sites directory is strongly confirmed by the very low level of overlap with other directories, especially from Voila Pages Perso to other directories, even though Voila Pages Perso is the smallest directory of all.
275

<table>
<thead>
<tr>
<th>Shares n % of its URL with</th>
<th>MSN</th>
<th>Nomade</th>
<th>Open Directory</th>
<th>Voila</th>
<th>Voila PP</th>
<th>Yahoo</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSN</td>
<td>100.0%</td>
<td>33.8%</td>
<td>13.2%</td>
<td>22.7%</td>
<td>1.5%</td>
<td>35.7%</td>
</tr>
<tr>
<td>Nomade</td>
<td>11.2%</td>
<td>100.0%</td>
<td>9.3%</td>
<td>19.7%</td>
<td>3.4%</td>
<td>30.4%</td>
</tr>
<tr>
<td>Open Directory</td>
<td>18.8%</td>
<td>39.9%</td>
<td>100.0%</td>
<td>22.9%</td>
<td>2.4%</td>
<td>33.6%</td>
</tr>
<tr>
<td>Voila</td>
<td>17.6%</td>
<td>45.8%</td>
<td>12.4%</td>
<td>100.0%</td>
<td>3.8%</td>
<td>37.7%</td>
</tr>
<tr>
<td>Voila PP</td>
<td>2.4%</td>
<td>16.6%</td>
<td>2.7%</td>
<td>8.0%</td>
<td>100.0%</td>
<td>8.7%</td>
</tr>
<tr>
<td>Yahoo</td>
<td>1.4%</td>
<td>39.5%</td>
<td>10.2%</td>
<td>21.1%</td>
<td>2.3%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Key: 11.2% of URLs of Nomade are also indexed by MSN, while 33.8% of the URLs of MSN are in the Nomade base.

Table 4 - Overlaps between directories

The mean overlap rate between the different directories is 18.1% and 22.8% if we exclude the very specific Voila Pages Perso. As in the case of Voila Pages Perso, size does not seem to be the decisive factor in overlaps between directories. For example, MSN and Open Directory, both small, share less than a third of their URLs, on average, with other directories up to four times bigger – an overlap equivalent to that between Nomade and Yahoo, the two biggest. It thus appears that each directory indexes sites peculiar to it. This is confirmed by the proportion, in each directory, of URLs indexed in no other directory (see Table 5).

<table>
<thead>
<tr>
<th>Directory</th>
<th>Number of sites indexed</th>
<th>Number of sites peculiar to the directory</th>
<th>Percentage of sites peculiar to the directory</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSN</td>
<td>46,22</td>
<td>21,174</td>
<td>45.8%</td>
</tr>
<tr>
<td>Nomade</td>
<td>139,0</td>
<td>74,089</td>
<td>53.3%</td>
</tr>
<tr>
<td>Open Directory</td>
<td>32,49</td>
<td>12,292</td>
<td>37.8%</td>
</tr>
<tr>
<td>Voila</td>
<td>59,80</td>
<td>21,482</td>
<td>35.9%</td>
</tr>
<tr>
<td>Voila PP</td>
<td>28,33</td>
<td>20,614</td>
<td>72.8%</td>
</tr>
<tr>
<td>Yahoo</td>
<td>107,0</td>
<td>48,899</td>
<td>45.7%</td>
</tr>
</tbody>
</table>

Table 5 - Proportion of indexed URLs peculiar to each directory

With the exception of Voila Pages Perso, which has a particular content (73% of URLs peculiar to it), we note that Nomade, the biggest directory, is also the one with the most particular characteristics (53.3%). This result was expected. Less predictable were the rates of particularity of MSN (45.8% of URLs peculiar to
it), which is three times smaller than Nomade, and Yahoo, which has few particularities despite its size. It thus seems that there is a twofold effect contributing to the particularity of directories: their size, which statistically increases their chances of indexing sites that others exclude, and their content, determined by their choice of indexed sites.

6. Web directories have strong identities

In this part we examined the particularity of each directory in terms of content and style. We first wished to establish whether, on a given topic (e.g. art, sport or politics), different directories have marked differences in choice of content and, if so, to what extent? We excluded Voila Pages Perso from this analysis, because of its intrinsic particularity, and retained only the five general-interest directories.

1. We qualified the content of the directories on the basis of the short descriptions they give of indexed Web sites on a given theme. The following method was applied:

   2. We first chose topics present at the first level, for the five directories studied;

   We then extracted, for each directory and for all the sites classified under the chosen topic, the directories' descriptions of those sites;

   3. The corpus thus constituted was processed with a lexicometric tool, Alceste (Reinert 1993). This enabled us to identify the specific vocabulary used by each directory to describe sites on a given topic.

The first topic selected was "Art and Humanities" (Art et culture) present at the top level of the five directories. At this level MSN focuses on North American museums (specific vocabulary: "US", "Canada", "Montreal", "New York", etc.), Open Directory gives priority to downloading of music ("MP3", "server", "free", etc.), Nomade emphasizes the accessibility of art ("consult", "invite", "share", etc.), Voila is more business-oriented ("bargains", "shopping", "catalogue", "order", etc.) and Yahoo seems to be more eclectic, with a slight preference for cinema ("scenario", "critique", "synopsis", etc.).

The same analysis was applied to the category "Business and economy" (Commerce et économie), also found at the first level of the five general-interest directories. Here again, MSN focuses largely on North American sites, with particular emphasis on financial topics ("financial", "bank", "investment", "shares", etc.). Nomade seems to be oriented more towards the tourism industry ("visit", "restaurant", "rating", "hotel", etc.). Yahoo and Open Directory tend to favour international trade and the manufacturing sector ("truck", "rubber", "development", "international", etc.). Lastly, Voila's position appears to be less specific although it has a slight tendency towards local business (place names are over-represented).

In addition to this, we performed a part-of-speech tagging of all the site descriptions for each Web directory in order to see its stylistic specificities. We thus analyzed the distribution of main POS categories and of verbs and pronouns person/number for each directory (see Table 6). This analysis raised an opposition between two presentational attitudes: on the one hand, directories as "helpful guides", with an over-representation of verbs and the 2nd person plural (typically: "You will find on this site...") : Nomade, Voila, Voila Pages Perso; on the other hand, directories as "neutral information relay", with very few verbs and a
dominance of nouns, adjectives and the 3rd person singular: MSN, Yahoo and Open Directory.

7. Discussion

Most earlier studies on Web directories were undertaken by researchers and specialists in library science and documentation (Bertonèche, 2001) (Chan, Xia et al., 1999) (Van der Walt, 1998) (Vizine-Goetz, 1996). Other studies have focused on directories as systems of reference classification, and have used them as a resource for automatic classification of documents (Mladenic, 1998), (Labrou and Finin, 1999). Two studies in particular attracted our attention in so far as they present a comparison of several Web directories (Bertonèche, 2001), (Van der Walt, 1998). In both cases the approach is purely qualitative and the findings are very close to our own, i.e. wide diversity, even heterogeneity, among the directories studied.

Our study has the particularity of combining a qualitative approach ("manual" analysis of directories and critical analysis of their modes of organization), and a quantitative approach based on statistical and formal processing: exploitation of the textual content of site descriptions in directories and of the structure of the directory as a graph.

<table>
<thead>
<tr>
<th>POS main cat.</th>
<th>MSN</th>
<th>Nomade</th>
<th>Open Dir.</th>
<th>Voila</th>
<th>Voila PP</th>
<th>Yahoo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nouns</td>
<td>78.5%</td>
<td>66.9%</td>
<td>73.8%</td>
<td>74.0%</td>
<td>70.6%</td>
<td>78.9%</td>
</tr>
<tr>
<td>Verbs</td>
<td>5.1%</td>
<td>16.4%</td>
<td>8.4%</td>
<td>9.8%</td>
<td>12.8%</td>
<td>5.6%</td>
</tr>
<tr>
<td>Adjectives</td>
<td>14.9%</td>
<td>12.3%</td>
<td>15.1%</td>
<td>13.2%</td>
<td>12.2%</td>
<td>14.2%</td>
</tr>
<tr>
<td>Adverbs</td>
<td>1.5%</td>
<td>4.4%</td>
<td>2.7%</td>
<td>3.0%</td>
<td>4.4%</td>
<td>1.3%</td>
</tr>
<tr>
<td>1 SG</td>
<td>1.3%</td>
<td>0.9%</td>
<td>0.9%</td>
<td>0.2%</td>
<td>8.8%</td>
<td>1.0%</td>
</tr>
<tr>
<td>2 SG</td>
<td>9.0%</td>
<td>0.9%</td>
<td>9.6%</td>
<td>2.3%</td>
<td>3.7%</td>
<td>26.0%</td>
</tr>
<tr>
<td>3 SG</td>
<td>79.7%</td>
<td>51.2%</td>
<td>60.6%</td>
<td>54.3%</td>
<td>46.8%</td>
<td>57.2%</td>
</tr>
<tr>
<td>1 PL</td>
<td>1.0%</td>
<td>1.3%</td>
<td>2.9%</td>
<td>0.2%</td>
<td>3.6%</td>
<td>0.9%</td>
</tr>
<tr>
<td>2 PL</td>
<td>2.2%</td>
<td>38.3%</td>
<td>15.5%</td>
<td>34.6%</td>
<td>30.1%</td>
<td>3.5%</td>
</tr>
<tr>
<td>3 PL</td>
<td>6.7%</td>
<td>7.4%</td>
<td>10.5%</td>
<td>8.5%</td>
<td>7.0%</td>
<td>11.5%</td>
</tr>
</tbody>
</table>

Table 6. Distribution of main POS categories and person/number for each directory

Specialists in cataloguing and library science recommend taking into account the different classificatory and methodological theories developed by their
disciplines, and applying them to the cataloguing of Internet sites. Our view is that this classificatory approach cannot be transposed as such to the

Internet world, for at least two reasons. The first is that the Web is not an encyclopedia of knowledge and is not comparable to a library. It offers content, services and, more generally, resources of all kinds, with a very wide variety of topics and quality. The second reason lies in the diversity of contexts in which the Internet is used, and of user profiles and needs. This is a very different situation from readers in a library or other documentary resource center, whose needs can be determined a priori and whose profiles have been defined on the basis of a long history of practice.

We agree with library science specialists on the need for more rigor in the construction of Web directories. We also think that general-interest Web directories, like those that currently exist, have possibly reached their limits. The management of a large number of sites and categories seems to pose problems manifested in a lack of coherence. Two areas appear to be emerging for further development of Web directories suited to current trends in their use: "regional" directories, specialized by sector of activity, geographic or cultural domain, etc., and systems based on collaborative evaluation and cataloguing of web sites within "interest communities".

References
Commercial Websites and the Use of Classification Schemes: The Case of Amazon.com

Abstract: The structure and use of the classification for books on the amazon.com website are described and analyzed. The contents of this very large website are changing constantly and the access mechanisms have the main purpose of enabling searchers to find books for purchase. This includes finding books the searcher knows about at the start of the search, as well as those that might present themselves in the course of searching and that are related in some way. Underlying the many access paths to books is a classification scheme comprising a rich network of terms in an enumerative and multihierarchical structure.

1. Introduction.

In a previous paper (Kwanik & Liu, 2000) we reported on a classification scheme used on the commercial auction website, eBay.com. The aim was to explore the nature and use of classifications in web environments where:

- the contents of the sites are in constant flux;
- the user population is unknown, or if it is known, we can assume it is diverse;
- it is desirable that the classification be very simple and straightforward so that all levels of users can learn it; and where
- the web environment provides multiple access routes and easy, flexible, and complex representations, so a classification system with requisite richness is desirable so that these functions can be supported.

In our previous work we chose eBay.com because it was an example of a dynamic, large, messy, and inconsistent, but surprisingly robust classification. The aim of the classification in eBay is not so much to provide access to specific items, but rather to maximize the chances of a person coming across an item (and thus, bidding on it). In this paper, we will extend our inquiry of the use of classifications in commercial websites to the scheme implemented for books in amazon.com. This large, commercial site is similar to eBay.com in its dynamism, and also in its provision of many access points and multiple, flexible routes for finding items — indeed for items being “found” even without searching. It is also similar in that the aim of the site and its classification is to encourage and enable people to purchase books now, in the present, and not necessarily to provide long-term, predictable and enduring future access routes to them, as in a library.

The amazon.com classification is different from eBay’s, however, in that it is not the users who classify the items from a list of possible headings. Rather, the classification terms as well as the “thesaurus” from which terms are drawn are designed by amazon.com. This presents some additional opportunities for extending
our knowledge of how such classifications are applied, and what role they play in the access to items.

As in the previous study, we approach the analysis of amazon.com's classification scheme using the following criteria:

- **scope of the classification** — what does it cover?
- **vocabulary** — level, consistency
- **structure** — the explicit and implicit relationships among entities
- **granularity (scale)** — the level of specificity
- **expressiveness** — how well does it reflect the domain it classifies?
- **hospitality** — how well does it accommodate new concepts?
- **browsability** — how well does it support and facilitate exploration of the domain?
- **usability** — how easy is it to use?
- **coherence** — how well does it "hang together"?
- **consistency** — is it predictable?
- **exhaustivity** — how completely does it cover the domain?

2. The amazon.com classification.

The classification scheme of "subjects" on the amazon.com website is a richly connected network of interrelated terms representing not only the topics of books, but also many other aspects. Thus the scope of the scheme encompasses both content and form. Here are the top categories as presented at the first level of the "Browse Subjects" feature offered when a book search is initiated.

Each of these categories can be further expanded. Categories at the top level comprise quite a diverse collection of entities. Some are topics (*History, Sports*); some are genres (*Science Fiction, Romance*); some reflect a perspective (*Christian Books, Gay & Lesbian Books*); some refer to format (*Audiobooks, e-Books, Large Print*). The *Teens* category could be about or for teens, while the *Today's Deals* category stands apart as unique in emphasizing price over topic, form, or audience.

<table>
<thead>
<tr>
<th>Arts &amp; Photography</th>
<th>Cooking, Food &amp; Wine</th>
<th>Large Print</th>
<th>Religion &amp; Spirituality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audiobooks</td>
<td>e-Books</td>
<td>Literature &amp; Fiction</td>
<td>Romance</td>
</tr>
<tr>
<td>Audio Downloads</td>
<td>Entertainment</td>
<td>Mystery &amp; Thrillers</td>
<td>Science</td>
</tr>
<tr>
<td>Biographies &amp; Memoirs</td>
<td>Espanol</td>
<td>Nonfiction</td>
<td>Science Fiction &amp; Fantasy</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sports</td>
</tr>
<tr>
<td>Business &amp; Investing</td>
<td>Gay &amp; Lesbian</td>
<td>Outdoors &amp; Nature</td>
<td>Teens</td>
</tr>
<tr>
<td>Children's Books</td>
<td>Health, Mind &amp; Body</td>
<td>Parenting &amp; Families</td>
<td>Today's Deals</td>
</tr>
<tr>
<td>Christian Books</td>
<td>History</td>
<td>Professional &amp; Technical</td>
<td>in Books</td>
</tr>
<tr>
<td>Computers &amp; Internet</td>
<td>Home &amp; Garden</td>
<td>Reference</td>
<td>Travel</td>
</tr>
</tbody>
</table>

Further, the top classes are not at the same level of granularity. For instance *Literature & Fiction*, when expanded, includes many other categories, including some of the top categories. Nor is the classification mutually exclusive. A given
category can be found under several of the top categories, as well as under several subcategories, at different levels of specificity.

It is not possible to initiate a search in amazon.com through the subject categories. Instead a user starts by entering a term in a search box and is then presented with some top picks, and at that time is also given some related broad categories for additional searching. Let's say one types in "CAD" (Computer Assisted Design). Several suggested books are presented, and when one of them is clicked, for instance, *AutoCAD 2000: No Experience Required*, the details of information pertaining to that book are presented. Attached to the description, at the bottom of the page, are the various subjects associated with the book as assigned by amazon.com. At the same time, at the top of the page, one of the clickable tabs is "Browse Subjects." Expanding the class *Arts & Photography*, the subject path to this book is as follows:

A given book, such as the one in this example, may appear under several different categories, and even under different branches of the tree. The following classes are also suggested. Each of these takes the searcher on a different but often overlapping path.

Engineering>Mechanical>Drafting & Mechanical Drawing
Computers & Internet> Graphics & Illustration> General
Computers & Internet> Web Development> HTML, Graphics, & Design> Web Graphics
Computers & Internet> Graphics & Illustration> CAD> AutoCAD

---

**Accounting & Finance**
**Architecture**
**Business Management**
**Education**
**Engineering**
**Law**
**Medical**
**Professional Science**

**Professional & Technical**

**Arts & Photography**
**Architecture**
**Art**
**Art Instruction**
**Artists, A-Z**
**Fashion**
**Graphic Design**
**Performing Arts**
**Photography**
**Today's Deals**

**General**
**Architect's, A-Z**
**Architectural Standards**
**Building Types & Styles**
**Criticism**
**Drawing & Modelling**
**Historic Preservation**
**History**
**Interior Design**
**International**
**Landscape**
**Materials**
**Project Planning**
**Reference**
**Study & Teaching**
**Urban & Land Use Planning**

**Engineering**
**Mechanical**
**Drafting & Mechanical Drawing**

**Computers & Internet**
**Graphics & Illustration**
**General**

**Computers & Internet**
**Web Development**
**HTML, Graphics, & Design**
**Web Graphics**

**Computers & Internet**
**Graphics & Illustration**
**CAD**
**AutoCAD**
Terms are repeated, sometimes in a slightly different form, in a variety of positions in the classification. For instance, the term Architecture appears under Arts & Photography, but also as follows under the category Professional & Technical

3. Expressiveness

Expressiveness refers to how well the terms in a classification describe and reflect the domain being classified. In the case of amazon.com's classification, the sheer number of terms assigned to any one book ensures a high level of descriptive power. Thus a book on CAD can be found whether the user construes it as software, architectural technique, or graphics. Put another way, the structure of the amazon.com scheme allows for a faceted approach, without necessarily building a faceted structure. That is, the classification achieves a multi-perspective view by redundancy of paths rather than by attention to the expressiveness of any given part of the scheme. In case any one category does not work, or is not found, another is readily available. This also ensures that something is found, regardless of the search specifications—a feature that, of course, enhances the possibility of making a sale.

Another measure of expressiveness is the ability of the classification to accommodate every entity within it. Normally, the presence of a "General" or "Miscellaneous" category might be a clue that the classificatory scheme is insufficiently developed. In the case of amazon.com, however, the use of General is merely additional insurance that a book will be found. For example, the book, _Heaven Help Us: The Worrier’s Guide to the Patron Saints_, by Alice and Clare LaPlante, is assigned the following categories:

- Religion & Spirituality>Christianity>Catholicism>Inspirational
- Religion & Spirituality>Christianity>Catholicism>Saints
- Religion & Spirituality>General
- Religion & Spirituality>Spirituality>Prayer
- Biographies & Memoirs>General

The use of General does not imply that there is not a specific enough category for this book, but rather that in case a user does not think of it, or if he or she is merely browsing in a more general category, the book will be encountered.

4. Consistency, Coherence, and Usability

As already mentioned, the amazon.com scheme is not consistent in terms of scale, terminology, or structure. Terms are nested, arranged, and presented without much attention to overall coherence. Furthermore, there is no overall view of the categories and relationships among them. The schematic representation of the classification is not available to users, nor is there any one place where a user can learn about all the terms available (except for the list of top terms on the opening page of the _Browse Subjects_ tab). The categories are accessed piecemeal as one goes along. Each term is clickable, and leads the searcher either to other terms or a selection of books.

This messiness and lack of mapping and predictability, however, serves the user well in one respect: it removes the necessity for the user to follow a rational
and orderly path. At every click, a new set of options in terms of books, subjects, related objects, and so forth are presented. The user merely has to choose. The system is designed for maximum connectivity between all the parts.

The negative aspect of this is that the user never has a sense of "place." Multiple paths seem to lead to the same spot, and there is a feeling of "having been there before." Nor does the user know whether a search is exhaustively covering the topic. He or she must rely entirely on the site's assignments of categories. The confusion is mitigated somewhat by the availability for any given search of a list of subject categories that have been visited. The user can thus keep track of where he or she has been.

5. Subjects for browsing and subjects for searching.

Without explicitly saying so, the amazon.com site provides subject terms for browsing, and subject terms for direct searching. Here's an example: The book Justice Hall by Laurie R. King is a mystery that features a fictitious young woman named Mary Russell who becomes involved with another fictitious character (from a different series of mysteries), Sherlock Holmes. Here are the subject categories amazon.com suggests for browsing (presumably to find books like this one):

- Mystery & Thrillers>Mystery>British Detectives
- Mystery & Thrillers>Mystery>Historical
- Mystery & Thrillers>Mystery>Women Sleuths
- Mystery & Thrillers>General
- Mystery & Thrillers>Authors, A-Z>(K)>King, Laurie

In addition to these, and directly below this list one finds the following:

**Search for books by subject:**
- Russell, Mary (Fictitious character)
- Fiction
- Holmes, Sherlock (Fictitious character)
- Women detectives
- England
- Fiction – Mystery/Detective
- Mystery & Detective – Traditional British
- Mystery & Detective – Women Sleuths

It is not clear, nor is it explained anywhere on the site, what these subjects represent and how they're different from the browsing terms. Some seem to be from the classification scheme described above. Some seem to be more general (e.g., England, or Fiction), and some very specific (e.g., Holmes, Sherlock). The terminology is sometimes isomorphic with that of the classification scheme, and sometimes different. What's the difference between Mystery & Detective and Mystery & Thriller? It's difficult to tell. When clicked, these terms lead to lists of books, rather than to an expanded list of subjects. In other words, this is the mechanism for direct searching.
6. Comparison of amazon.com's classification with traditional bibliographic classifications

Dewey Decimal Classification (DDC) and the Library of Congress Classification (LCC) are traditional classification schemes for books used in the United States and elsewhere. These two classification schemes provide interesting comparisons with amazon.com because each emerged from somewhat different paradigms of classification design. DDC is a classification based on a model of knowledge, as reflected in academic disciplines and traditional disciplinary boundaries. LCC, on the other hand, emerged as a pragmatic way of organizing an existing collection, the actual books in the library that served the U.S. Congress. Over time, DDC and LCC have blurred the line between being “top-down” or “bottom-up” classifications, but the distinction that defined the core of these two schemes continues to inform many crucial decisions in classification design and knowledge representation: do we classify based on enduring consensual (and often academic) models of how things “go,” or do we classify based on popular and perhaps ephemeral models that may change, but which have the potential for being truer in some ways to the users’ perspective on the world?

The classification in amazon.com is an interesting blend of traditional approaches to classification and a creative application of the multiple access points enabled by web technology. For example, traditional classifications provide a conceptual ordering that has as one of its goals the presentation of any given concept in the context of its superordinate, subordinate, and related concepts. In amazon.com’s scheme we find this context as well in the expandable subject terms (as in DDC) and the subject strings presented to the user. At the same time, there is no special care taken with ensuring that this context is conceptually coherent, stable, or consistent.

In general, amazon.com’s scheme can be viewed as more pragmatic and enumerative than as based on a model of knowledge. In this way it is more like LCC. There is a difference, though, because in LCC the main classes are mutually exclusive, while in amazon.com it seems that the main classes are designed to reflect some equal distribution of postings (rather than conceptual congruence). So, for instance, there is a top category for Mystery & Thrillers. This is on the same level as Literature & Fiction, a category that might include M & T. One could argue that for conceptual consistency, Mystery & Thrillers might better belong in the subcategory, Genre Fiction under Literature & Fiction. The number of postings to the M & T category, however, makes this a very popular category, and thus it appears as a class unto itself.

7. Conclusion.

The study of classification in dynamic commercial websites has implications for the rethinking of classification work in general. While amazon.com has a core mission of selling things, we are coming to understand that such sites have generated many parallel non-commercial uses as well. It is not uncommon for amazon.com to be used for bibliographic work and collection development, for instance.

The classification scheme in amazon.com has been implemented to maximize the chance of finding a given book, but also of finding related books. In this the goals are similar to traditional library-classification goals, except that the
design and implementation are different. The classification in amazon.com uses multiple access routes, a simple but redundant vocabulary, and not much attention to structural integrity in order to provide a robust and rich network of subjects. This comes at the price of sometimes being lost, and of a lack of an overall view. If we could emulate the flexibility of the commercial site and preserve the structural integrity of the traditional library approaches we would have classifications that not only reflect the collection, but also better reflect the multiple perspectives of users.

References.
Design, Development and Management of an Information Recovery System for an Internet Website: from Documentary Theory to Practice

Abstract: A real case study is shown, explaining in a timeline the whole process of design, development and evaluation of a search engine used as a navigational help tool for end users and clients on a content website, e-commerce driven. The nature of the website is a community website, which will determine the core design of the information service. This study will involve several steps, such as information recovery system analysis, comparative analysis of other commercial search engines, service design, functionalities and scope; software selection, design of the project, project management, future service administration and conclusions.

1. Introduction

Information retrieval systems tell us generally about new investigations, even new retrieval models, but it is not frequent enough to find in the literature explanations on concrete experiences, or case studies.

This paper intends to explain, based on a real case study, the whole process of design, development and exploitation of an internal information retrieval system, within a content website, oriented specially on vertical communities and e-commerce. The system will give service both users/clients and the enterprise. Thus, we will follow the basic steps needed to transform theoretical approaches onto praxis, and how to apply the classical documentary techniques to online selling.

The website Temalia.com is the example used in this case. It is a web portal with more than a hundred vertical communities hobbies driven, with such diverse thematics as hunting, literature, biology or action cinema. The commercial nature of the website holds the key to understand the nature of the project and its scope.

The main objective of this service is to supply the bare information necessities of multicultural users, with different geographical and linguistic realities, even inside the Spanish language as a "lingua franca", with others true intentions:

- To serve as a bridge or an initial door to many different contents and the communities themselves.
- To answer the questions of the users.
- To show at once the whole world of services which Temalia.com has.
- To help users to become clients.

Let's see the process of making this real step by step:
2. Information recovery system analysis

Every project has to be developed following this first question: "what this will be useful for?"

The information professional tends to try to develop a "perfect" system, or, better said, a perfect system for him. This would be great if we were talking of a typical inner service, designed for information professional exclusive use. In that case, would be easier for the professional to better performance search queries, gaining feedback and expertise on using it and learning the best way to exploit the tool. But this is not the case.

In a web information recovery system, the mediation between text language and users, natural query language will be at least transparent for the user. So we have to focus on user. Usability will be a priority, and free-text based search will be the way we let users search.

Plus, we have to answer other basic questions to determine how we will design the system: Why use a search engine? What for? For whom is designed? Which implications has it for the whole enterprise, the content management or the business processes?

3. Comparative analysis of other commercial search engines

Before deciding the tool we will use, or the front page design, we should take care of what is doing some other websites and what type of capabilities have some typical search engines, the ones who users all around the world prefer. Benchmarking is required.

We will obtain some interesting comparative indicators. We will be able to analyze their main focus, their strategy on serving public, their own original characteristics and how they affect the performance of the whole system.

We will find coincidences and differences between them, learning from them, in such a way that we will be able to gain competitiveness by means not only of quality, but matching specific necessities matching.

4. Service design. Functionalities and scope

We will first think about how we are going to answer the problem of the cultural diversity of our potential users. How we can show the contents, not only the information, to the final user and which kind?

We decide to offer on the results of a search query a little bit of the different types of content and services that we have, bringing to the user the first navigational option. This idea will affect the rest of the project.

We will specify then which kind of service we need for our project, and how the system architecture is going to be organized, depending on the performance necessities.

Finally, and not before, we will start to search for commercial search tools, thinking of which technical requisites we need to look for. This is crucial for the success of the project. Normally enterprises look first at the tool, and then at the project. To the contrary, what we need is first determine why and what for we are going to use a tool, and then look for the tool.
Fig 1. Benchmarking search engine's services and functionalities

Fig. 2. Initial version model of front-end presentation of answers in Temalia’s search engine.

5. Software selection

Several commercial and non-commercial tools are studied. In the business world, it is very important to distribute efforts, so if we don’t have a great technical
force, we should think of using other human professionals to help us in the development of the service.

Another study is made, comparing the capabilities of every tool, and how they can match with our necessities. We generate more comparative indicators, focused on:

- their technical characteristics
- flexibility to adapt to our main architecture system
- integration with our management content system and content formats
- the company that will develop the project for us, and the number and expertise of their professionals involved. The success and variety of other similar experiences is very important when we consider the tool and the people which we are going to work with.

- Relation price / quality of service
- The priority is not to select the best software itself, the ultimate solution on the market, but the one which best covers our necessities.

6. Design of the project

Once we know which tool will be used, and what company will develop the project with us, we study all details around the development process, such as technical requirements, programming effort, and material and human resources involved.

The point is to find the best way to reduce the time to market, because in e-business this factor is critical. The user needs to find what we have got, and the company needs to offer the products which best match the user necessities, and time is money.

7. Project management

A schedule of actions and people involved is built-up, and shared between them, in order to have everybody aware of the others' responsibilities.

The steps will include installation of the software, adaptation to the architecture system, adaptation to the content management system, integration with diverse software and creation of front-end templates.

Several professionals collaborate in this interdisciplinary exercise: web designers, usability experts, information professionals, technicians, programmers, content managers, marketing and project directing. Periodical briefing meetings are scheduled, in a state-of-the-art way of working till the pre-fixed objective.

8. Service administration

After being first used in an beta mode and its kick off, a constant period of maintenance is commenced. If worthwhile, a professional should feed the inside thesauri tool with new keywords, building up semantic relations which better help the information recovery continuous feedback.

Example: We have a section of cinema. The film originally entitled in English "The Sound of Music", in Spain was translated into 'Sonrisas y lágrimas', but in Chile "La novicia novata". This tells us about the immense variety of misspellings, cultural and geographical variations, etc. The service should try to
fight this kind of problem, and one of the better ways is enabling and integrating a thesaurus tool to build all kinds of semantic or commercial relationships.

Search logs are one of the best and least tools used to study CBA (Customer Behavior Analytics) since search behaviour offers an unique opportunity to analyze proactive queries about contents, hobbies and products, which will help us to coordinate the products placement between the contents, and the selection of new offers with more probabilities of selling success.

Plus, we will take care of the efficiency and effectiveness of the system. The system will be designed to enhance user satisfaction, helping users to navigate, even if we do not have what user is looking for.

Conclusions

After a certain period of time, the whole project will have to be re-evaluated, because in e-businesses, mobility and adaptivity are common coin. With user feedback and changes of designs or necessities of the product (defining 'product' as the whole website itself) we could think again about the rate of success of the service.

Learning that nothing on Internet is static, we will have the ability to change before, better and rapidly. A system flexible then, with a developing application, will be needed, in order to quickly change what any new project design demands.

References


José Luís Herrera Morillas and Mª del Rosario Fernández Falero
University of Extremadura, Faculty of Librarianship Studies and Documentation, Badajoz, Spain

Information and Resources about Bibliographic Heritage on the Web Sites of the Spanish Universities

Abstract: The main objective of this paper is to study and compile the information and facilities provided by Spanish university libraries in relation to library patrimony. By patrimony we understand old, historic works. The methodology consisted of an initial analysis of the websites of the various Spanish university libraries, focusing on those areas of knowledge and resources of interest to us; we then proceeded to study the data collected.

1. Introduction

The university library patrimony, both quantitatively and qualitatively, is no any doubt an important part of the Spanish library patrimony.¹

New advances in technology are revolutionising the library panorama. In Spain, University libraries have taken the lead in making use of the possibilities provided by the Web. In this paper, we first conducted a tour of the websites of Spanish universities, collecting the information and facilities relating to library patrimony available on the Internet.²

Secondly, we classified and quantified the data obtained, and added some considerations.

2. Review of the patrimony of Spanish university libraries on the Internet

2.1 Barcelona University Library.
http://www.bib.ub.es/bub/bub.htm

This library has a section dedicated to Reserve Stock <http://www.bib.ub.es/www7/7res1.htm> which contains information from the old stock of this University. It includes references with data on the different sections of the old title collection, and other sections such as: “Historic Archives”; “Showcase” (“Vitrina”), a periodical display of works from the collection; “Recently Found Works of Value”; “Restoration and Conservation”; “Digitisation Projects” (internal and external); Virtual Expositions and Museums (internal and external); “Bibliophilia” (related addresses on the Net); “Special Collections” (a record of modern and old titles of BU, the Botanic Institute of Barcelona, Ramón Llul in the BUL, Old Book Networks); and “Other Useful Ways of Finding Information on Old Titles”.

The following expositions were organised by the BUL, and are listed under “Virtual Expositions” http://www.bib.ub.es/www7/7res2.htm Exposition of Old Titles: Dermatology in the Library Reserve Collection; Gynaecology and Obstetrics in the Library Reserve Collection; and Bibliographic Treasures in the Barcelona University Library.
The website of this Library also gives access to its own catalogue of old titles: *Catalogue of old titles: documents published before 1820* http://eclipsi.biblub.es/virtua2/catalog/index.html

Other projects of note are the *Makes of printing presses* data base http://eclipse.bib.ub.es/imp/impcat.htm, which permits the visualisation of images of some of the makes of printing presses included in the catalogue of old stock, and the digitisation project called *Fons grewe (The Grewe titles)* http://www.bib.ub.es/grewe/grewe.htm, which is a selection of digitised books from the collection on cookery and gastronomy. Access can be gained to the images through the catalogue on old stock or through the BUL website. For access to the digitised books through the catalogue, a specific label is included in the bibliographic register; the second option – access through the website – includes a list of digitised books in alphabetical order of authors, separated by century of publication, and with a brief bibliographical description.

2.2 Castilla-La Mancha University Library http://www.uclm.es/

This library has a virtual exposition called *The Quixote Library* http://www.biblioteca.uclm.es/cervantes/cervantes.htm which comprises 47 editions and studies on the Quixote and Cervantes, and 14 printed illustrations from the Cumplido and Gregorio Prieto editions of the Quixote. The main aim is to provide a small representation of specialised titles on the Quixote and Cervantes from the libraries of the four campuses of the University (Albacete, Ciudad Real, Cuenca and Toledo), and to present the latest acquisition on this subject, i.e. the private collection of José Esteban.

As from 27 October 2000 the Carmen Bravo-Villasante Bibliographic Collection http://www.biblioteca.uclm.es/Cepli/biblioteca.htm can be viewed in the General Library of the Cuenca Campus. This collection includes, amongst others, works edited between 1728 and 1939.

2.3 Madrid Complutense University Library http://www.ucm.es/BUCM/

In addition to the *Historic Collection Catalogue* http://cisne.sim.ucm.es/screens/opacmenu_spi_sl.html, found in the automated catalogue “Cisne” of the MCU, the facilities available relating to old stock are centralised in the recently opened “Marqués de Valdecilla” Historical Library. It can be accessed through the innovations section http://www.ucm.es/BUCM/frames09.htm of the MCUL website and contains information on stock, services and specialised information; access is also provided to virtual expositions (organised by the Library), sale of publications and the Dioscórides Digital Library.

By entering the information section of the general index of this website, we can also gain access to the above-mentioned virtual expositions, which include the following:

- *Athanasius Kircher and 17th Century Science; History of the Book Through the Complutense University Collections; Old Mathematics Books; The Year 2000, World Mathematics Year; Geography Books of the Complutense University: from Ancient Times to the 18th Century; Three Centuries of Medical Knowledge in Madrid Complutense University, 15th to 18th Centuries; and The Complutense University Through its Books: 500 years of the Cisnerian Bull.*

A digitisation project remarkable for its ambitious objectives is the *Dioscórides Digital Library*, related to the Project of the same name <http://www.ucm.es/BUCM/diosc/00.htm>. This is a digital library on Health
Sciences, which came into being as a result of a project to digitise and store on optical-magnetic file the historical biomedical collection of titles from the 15th to the 18th centuries contained in the “Marqués de Valdecilla” Historical Library. The digitised catalogue of old titles contains a bibliographical description of the digitised books together with an image of the cover or some representative pages of each work.

2.4 Córdoba University Library
http://www.uco.es/webuco/buc/

The Córdoba University website, in the section on the Rabanales Campus Library, contains a subsection on old and historic titles from this University <http://www.uco.es/webuco/buc/bcr/fondoa.htm>. Under the epigraph “Organisation” of this website, as an annex to the Library Regulations, one can find Rules of Use and Conservation of Old and Historic Titles of the Córdoba University Library
http://www.uco.es/webuco/buc/buco/reglamen.htm#ANEXO.

2.5 Granada University Library
http://www.ugr.es/%7Ebiblio/

From the website of this Library access can be gained to a catalogue specifically dedicated to old titles http://fondant.ugr.es/ajibe/consulta.html. This catalogue is linked to the Aljibe database, which contains at present 13,038 entries. Included are: a catalogue of the rare incunable manuscripts which are kept in the Library safe, works printed in Latin from the 16th and 17th centuries, works printed in Granada from the 16th to the 19th centuries, and an undetermined number of old printed works which do not belong to either of the above categories and which have been included for other reasons (microfilming...). One of the most interesting qualities of the Aljibe database is that it allows the visualisation of 5,545 images in JPG format corresponding to covers, colophons and representative illustrations of the catalogued documents.

The information on old stock of this Library is completed with “Historical Review”
http://fondant.ugr.es/ajibe/historia.html which includes sections on the history of the Library, the composition of the historic collection, the Aljibe database, and manuscripts and printed works of special interest.

2.6 Huelva University Library
http://www2.uhu.es/biblioteca/

Available in this website are Rules for the Old Titles Room http://www2.uhu.es/biblioteca/norma2.html which provide guidelines for consulting and handling the collection of old books, facsimiles, legacies, theses and publication stock of Huelva University.

2.7 Jaén University Library
http://www.ujaen.es/serv/biblio/

Under the normal entry for catalogues, http://www.ujaen.es/serv/biblio/catalogos/catalogos.html in the section devoted to subcatalogues, all the special collections can be found, among them the old title collection.
2.8 La Laguna University Library http://papyrus.bbtk.ull.es/index.htm


2.9 Málaga University Library
http://www.uma.es/servicios/biblioteca/default.htm


2.10 Oviedo University Library
http://librivision.uniovo.es/web/

From the website of this library we selected “General Information” and then the heading “About the OUL”, which contains information on the old stock collection of the Oviedo University Central Library.

Another section under the General Information entry provided us with links to the different University libraries, one of which is the library of the Feijoo Institute of 18th Century Studies. Here we found a rich collection of monographs and specialised periodicals from the 18th century, and a remarkable collection of old printed works and manuscripts.

2.11 Catalunya Polytechnic University Library http://bibliotecnia.upc.es/

This University provides general information on its Historic Science and Technology Collection http://bibliotecnia.upc.es/bib240/Fonsanti.htm and on two virtual expositions: Thermal energy and its applications: historic elements: 18th, 19th and 20th centuries http://bibliotecnia.unc.es/bib240/expo_xviii.htm, and New acquisitions of technical books from the 18th and 19th centuries http://bibliotecnia.unc.es/bib240/expo_xviii.htm

2.12 Pontificia de Comillas University Library
http://130.206.67.16/portales/Biblioteca/default.htm

The website of this University, under the entry “History of the Library”, contains data on the old collections preserved in this Centre, of which the collection of scholastic works is worthy of note.
2.13 Pompeu Fabra University Library <http://www.upf.es/>

Information on "Reserve Stock" is available on: http://www.upf.es/bib/cont4.htm?d=/bib/actualit/visita.htm. Worthy of note are the important collection of the works of Pompeu Fabra, and the incunabulum *Catalonia Constitutions* (1495).


In the website of the Library of the Reserve Collection of Jurisprudence Titles

http://www.urv.es/sgenerals/biblioteca/Bib_Juridiques/index.html, under the section dedicated to documentary and bibliographic titles, there is an entry for "reserve stock titles" (libros del fondo de reserva) which contains a list of the bibliographical references of these works linked to catalogographic records.

The website of the Education and Psychology Library

http://www.urv.es/sgenerals/biblioteca/Bib_Educa/index.html has an entry in its index for "stock" which is linked to a section on old titles and information about this collection.

2.15 Salamanca University Library
http://sabus.usal.es/bibliotecas.htm

This library provides online access to *Catalogue of historic titles*

http://sabu.usal.es/catalogo.htm

2.16. Seville University Library
http://bib.us.es/bibliotecas/bgualasp

The Webster of the General Library is linked, under the heading "Old stock" http://bib.us.es/bg/fondoant.asp to an informative text on the history and content of the collections of manuscripts, incunabula, printed works from the 16th to the 18th centuries, the palatine library within the General Library, and related bibliography.

The website also contains an index with access to a text on *Rules for the Use and Conservation of the Old Stock Collection* <http://bib.us.es/bg/normafondo.asp>, a *Request for reproduction of old titles* application form <http://bib.us.es/bg/formulario.asp> which can be used to request information on the cost of acquiring reproductions of works from the old stock collection, and to *Internet Resources* <http://bib.us.es/recumate/fondoant.asp>, a list of subject-classified addresses related to old titles.

In addition, the Library offers, under the entry *Fame Catalogue*,

http://www.us.es/include.frameador.php?url=http://bib.us.es/ access to several catalogues, one of which is on works published before 1801. Also included here are automated records of publications from the 15th to the 18th centuries with digitised images.

2.17 Valladolid University Library
http://cigales.cpd.uva.es/biblioteca/principal.htm

The libraries section of the website of the University Library gives access to the Santa Cruz Historic Library. This is the General Section of the Old Stock Collection of the University, containing all the works published before 1835 as well as the Santa Cruz Library’s own collection.

2.18 Valencia-Estudi General University Library http://www.uv.es/~infobib/

In this University the old titles collection is in the Historic Library, which has its own website <http://www.uv.es/~bibhi/bibhistorica/> with information on human resources, stock, services, links of interest and news.
2.19 Zaragoza (Saragossa) University Library
http://wzar.unizar.es/doc/buz/bibliotecas/centros.html

This library has designed an exposition entitled *Exposition of Americanist Book Collections of the Zaragoza University General Library*
http://wzar.unizar.es/doc/buz/biblioam/catalogo.html

2.20 REBIUN (Spanish initials for University Libraries Network)

This library network has a working group dedicated to library patrimony
http://www.uma.es/rebiun/partimonioBibliografico.html with information on its main objectives. It includes a report on activities carried out in the year 2000, one of which is the exposition *Ex-Libris Universitatis*
http://busc.usc.es/exlibris/Paginas/segund.htm, providing details of Spain’s important library patrimony. The exhibits consist of 209 works from the 10th to the 18th centuries.

3. Results and conclusions

We analysed the websites of a total of 66 Spanish university libraries. Of these, 19 have their own facilities related to library patrimony. The overall number of facilities compiled is 53. For purposes of classification, we propose the following scheme, in which the number of facilities found for each type is indicated:

- Specialised libraries: 4
- Catalogues: 10
- Virtual expositions: 17
- General information on old book collections: 11
- Digitisation and reproduction projects: 5
- Specific rules for the use of old books: 4
- Compilation and access to external resources: 2

The libraries with most facilities are Madrid Complutense University (10 facilities), Barcelona University (8 facilities) and Seville University (6 facilities).

These data reveal that Spanish university libraries are making use of the Web to inform, diffuse and offer access to library patrimony.

Within these facilities, a distinction can be made between those focused exclusively on information and diffusion, and those giving access to the works by means of digitisation techniques. Outstanding examples are the Dioscórides Library (Madrid Complutense U.), the Aljibe databases (Granada U.) and the makes of printing presses (Barcelona U.), together with the wide range of virtual expositions.

We consider that these facilities are of great interest because they cater for both preservation and access: "the two objectives that in the age of paper were at loggerheads, can now be in consonance in the digital era".

The virtual expositions are the most numerous among the types of facility established, and of those compiled here the following are outstanding for their volume and coverage: *Ex-Libris Universitatis, Exposition of Americanist Book Collections of the Zaragoza University General Library* and *The Complutense University Through its Books: 500 years of the Cisnerian Bull*. The majority of these expositions are similar in structure: an introduction, a catalogue with images, a description of the documents and, in some cases, other elements such as bibliography, links to related facilities, credits, etc.

The most notable projects have been carried out by those libraries with the longest tradition and the richest collections of old books, i.e. Madrid Complutense,
Granada, Seville or Barcelona, which are "historic libraries in the forefront among the old university districts".

We believe that the principal importance of these initiatives is that collections of old works may now be released from their isolation and be made available not only to specialists but to all who wish to have access to them.

Notes
2. The data presented in this paper correspond to the most recent consultation of websites, carried out in February 2002.
Metadata Functionality for Semantic Web Integration

Abstract. We propose an extension of a mediator architecture. This extension is oriented to ontology-driven data integration. In our architecture ontologies are not managed by an external component or service, but are integrated in the mediation layer. This approach implies rethinking the mediator design, but at the same time provides advantages from a database perspective. Some of these advantages include the application of optimization and evaluation techniques that use and combine information from all abstraction levels (physical schema, logical schema and semantic information defined by ontology).

1. Introduction

Although the Web is probably the richest information repository in human history, users cannot specify what they want from it. Two major problems that arise in current search engines (Hefflin, 2001) are: a) polysemy, when the same word is used with different meanings; b) synonymy, when two different words have the same meaning. Polysemy causes irrelevant information retrieval. On the other hand, synonymy produces lost of useful documents. The lack of a capability to understand the context of the words and the relationships among required terms, explains many of the lost and false results produced by search engines.

The Semantic Web will bring structure to the meaningful content of Web pages, giving semantic relationships among terms and possibly avoiding the previous problems. Various proposals have appeared for meta-data representation and communication standards, and other services and tools that may eventually merge into the global Semantic Web (Berners-lee, 2001). Hopefully, in the next few years we will see the universal adoption of open standards for representation and sharing of meta-information. In this environment, software agents roaming from page to page can readily carry out sophisticated tasks for users (Berners-Lee, 2001).

In this context, ontologies can be seen as metadata that represent semantic of data; providing a knowledge domain standard vocabulary, like DTDs and XML Schema do. If its pages were so structured, the Web could be seen as a heterogeneous collection of autonomous databases. This suggests that techniques developed in the Database area could be useful. Database research mainly deals with efficient storage and retrieval and with powerful query languages.

The database community has been seriously disturbed with the recent Web technologies expansion. Particularly, two reports have produced special commotion in database field. The first one, the Asilomar report (Bernstein, 1998), postulated the new directives in databases tendencies, previewing the Web impact in this field. The second one, Breaking out the Box (Silberschatz, 1996, 1997), proposes how database community must transfer its technology to be introduced into Web technology. In this context we have broken out the database box and into its
autonomous functional components, and we are using these to reach a solution for the problem of heterogeneous service integration.

In recent years, mediators are responsible of integrating databases, using wrappers to translate the different data source information to a common model. They require a single schema for query processing. This means that possible local data source schema extension will not be available until the mediator and its wrappers are updated. Ontology-based systems use the semantics of data for data source integration. They must supply ontology query processing. So they do not need a single schema, giving dynamism to query processing.

2. Background

Wrappers were the first building block for Web integration. They act as interfaces to each data source, providing (semi) structure to non-structured sources or mapping the original data source structure to a common one. Unfortunately, it is very difficult to switch unstructured data into a specific schema. Issues related to wrapper design and implementation can be found in (Roth, 1997), (Sahuguet, 1999a; 1999b).

The knowledge about evaluating a query over multiple wrappers is encapsulated by mediators. The wrapper-mediator approach provides an interface to a group of (semi) structured data sources, combining their local schemas in a global one and integrating the information of local sources. So the views of the data that mediators offer are coherent, performing semantic reconciliation of the common data model representations carried out by the wrappers. One of the main problems in data integration is related to the maintaining of the integrity constraints. As this problem is not yet solved, mediators need to deal with the problem of evaluating consistent queries over possibly inconsistent sources with respect to the common schema constraints. Some good examples of the wrapper-mediator systems are AMOS (Fahl, 1993), TSIMMIS (Garcia, 1995), DISCO (Tomasic, 1995), GARLIC (Roth, 1996; Hass, 1997; 1999). Recently, many of these approaches have moved toward XML standard, like AMOS and TSIMMIS. On the other hand, MIX (Baru, 1999) (the successor to the TSIMMIS project) and MOCHA (Rodriguez, 2000) projects are initially XML-based.

The next level of abstraction on Web integration corresponds to ontology-based systems. Its main advantage over the mediators is their capacity for managing a priori unknown schemas. This is achieved by means of a mechanism that allows contents and query capabilities of the data source to be described declaratively. From the data perspective, Ontologies enrich the semantics of the schema, resolving synonymy and polysemy problems, viewed at the beginning of the paper.

The reader can find an excellent review of the state of the art in ontology engineering in (Corcho, 2001). Well-known environments for building ontologies are WebODE (Arpirez, 2001) and WebOnto (Domingue, 1998). The first one provides an API for ontology access (based on Prolog) and import/export utilities from/to diverse markup and ontology languages. WebOnto is a powerful collaborative environment focused on the ontology creation and navigation.

From the integration point of view, many studies have been and are still being developed using ontologies. At the moment this paper is being written we can remark on two main projects: Ariadne (Arens, 1996), (Baris, 2000) and Observer (Mena, 2000). Ariadne aims at the development of technologies and tools for
rapidly constructing intelligent agents to extract, query, and integrate data from Web sources. Observer uses different ontologies to represent information data sources. The user explicitly selects the ontology that will be used for query evaluation. The existence of mapping among ontologies allows the user to change the ontology initially selected.

The intended domain of our framework is closed to Observer. We designed an architectural framework for the integration of both heterogeneous online data sources into Web services and heterogeneous online services (Aldana, 2001b). This architectural framework enables the capability of an easy development and deployment of small and autonomous cooperating services. Furthermore, metadata information described in the services allows the cooperation and integration among them. In (Aldana, 2001a) we presented three different integration models: query delegation, service cooperation and semantic integration. Now our objective is to extend the framework meta-data component to support new concepts that emerge with the Semantic Web.

3. Integration Component

Traditionally, heterogeneous data source integration systems have been developed according to onion architecture. Following the nomenclature in (Stuckenschmidt, 2002), there are three layers: Syntax, Structural, and Semantic. Wrappers are in the syntax layer, which is the onion core. Mediators added a new layer, the structural one, covering wrappers. Currently, the most external layer is the semantic one that fits with ontology issues. Many of the developed systems match this architecture, or have been designed following it.

In Aldana (2001b) we presented a framework for the construction of services as Web component, mediators, which only reach the structural level. This framework follows some W3C consortium recommendations and latest Web technologies (Abitebould, 2000). XML is an appropriated standard for data interchange and nowadays is one of the most generalized ones. Our framework uses XML-Schemas and RDF-Schemas to represent schemas and specific information.

Our Framework helps the system developer to support them in the construction of services that integrate heterogeneous data sources, building a global schema for them (conceptual, logical and physical schemas). To construct such schema the developer has to know the data sources schemas and has to integrate them manually. The query processor accesses the data sources using the wrappers and combines their answer to present the final result to the user. This process is hidden to the user, who views the system as a single data source. The Framework architecture is shown in figure 1.

We replace the Framework Metadata Component (Aldana, 2001b) with a new one called Integration Component (IC) to support new Semantic Web emerging concepts. This new component achieves query processing over metadata information stored in the schemas. Besides, it allows the ontologies inclusion. Ontologies establish a joint terminology between members of a community of interest. These members can be human or automated agents. To represent a conceptualization, a representation language is need. However, for applications on the web it is important to have a language with a standardized syntax. Because XML emerges as the standard language for data interchange on the web, it is also
desirable to exchange ontologies using XML syntax. The integration component architecture is shown in figure 3.

![Figure 1: Framework Architecture](image)

Figure 1: Framework Architecture

![Figure 2: Metadata Component](image)

Figure 2: Metadata Component

![Figure 3: Integration Component Architecture](image)

Figure 3: Integration Component Architecture

The IC has the same metadata information as the metadata component plus the ontology information. This architecture allows performing two levels of query processing. The first one, the IC-processing, generates a complex schema from several heterogeneous data sources, using ontologies. The second one, the Query Processor Component Processing, performs specific application domain tasks. This two-level structure separates knowledge organization and the computing issues. Under this structural design changes in schema complexity, the use of alternative ontology algorithms, changes in user views, could be performed without affecting applications, being unnecessary modifications in their programming code.

Among other uses, we are especially interested in the ontology as a powerful mechanism for schema integration and schema enhancement. Given that the initial framework design was very close to a database organization, we have proposed an approximation to ontology-driven integration being consistent with our approach. On the other hand, ontology enriches the base model providing a set of valid models by applying axioms and rules. Since our interest is focused on how ontologies interact with the schema, ontology facilities will cover schema definitions, instead of being a new layer over the whole mediation service. Doing so we need to redesign the Metadata component of our architecture, but we gain some advantages.

We are more interested in the construction of mediators rather than in a specific mediation system, and in the study, development and trials of new
algorithms for Semantic Web Integration. Particularly we pretend to apply distributed deductive database evaluation and optimization techniques.

Extensions of Datalog (Ceri, 1990), a deductive database language, have been used as a formalism for Ontology definition. We think it could be very interesting to adapt many Datalog specific techniques to ontology-driven integration. Especially when mediation and ontology reside both into an integrated system, because it allows full access to mediator schema and ontology. For example, this allows the use of semantic query optimization techniques that use ontological constraints, schema constraints and physical level information (like access time statistics, the presence of indexation, etc.).

One of the most relevant aspects that the Web imposes in a database is the distribution aspect. Information distribution not only has effects on the efficiency of a database, but it affects the whole database management system, from data consistency and integrity issues to query processing. From a theoretical point of view, the Web favors the existence of distributed autonomous schemas. Under these circumstances, and considering ontologies as a specification model for advanced schemas, we are interested in data integration when ontologies, as a schema, are also distributed.

Our database research line was initially focused on deductive databases, with a special interest in distribution aspects. Recursion management in a distributed system is not an easy issue, and is more complicated in an environment like Internet. Questions related with cycle detection, evaluation synchronization and termination detection must be considered and studied carefully.

Finally, we think that an important factor for the growth of the Semantic Web is related with the users capability to query multiple ontologies that have been related in a decentralized way, using references or single relationships among its terms. This does not exclude but complements specifically constructed global ontologies. Currently, mediators can interoperate in a similar way as a distributed database does. So, we believe that is justified enough to fit ontologies together with the schema of the mediator, instead of locating them on top of it.

4. Conclusions and Future Work

We propose an extension of a mediator architecture dealing with ontology-driven data integration. In our architecture, ontology is not managed by an external component or service, but is integrated in the mediation layer.

The Framework Metadata Component (2001b) is replaced by a new component called the Integration Component. This new component achieves query processing over meta-information stored in the schemas. Besides, it allows ontology processing.

This approach implies rethinking the mediator design, but at the same time provides advantages from a database perspective. Particularly, it enables the application of optimization and evaluation techniques that use and combine information from all abstraction levels (physical schema, logical schema and semantic information defined by ontology), which is the focus of our current research.
References


Uta Priss  
School of Library and Information Science, Indiana University  
Bloomington, USA

Alternatives to the “Semantic Web”: Multi-Strategy Knowledge Representation

Abstract: This paper argues that the Semantic Web needs to incorporate both formal and associative structures (and possibly a multitude of other structures and strategies) to be successful. The arguments for this claim are based on an observation of successes and failures in the areas of artificial intelligence (AI) and natural language processing (NLP).

1. Introduction

The WWW provides numerous challenges for information and knowledge processing activities. Information may be available but not accessible or retrievable because of language barriers or insufficient search strategies. Data mining techniques may discover implicit information in explicit data but these techniques do not necessarily guarantee that the discovered information is relevant, significant and trustworthy. During the last several decades hundreds or thousands of computer and information scientists have developed probably thousands of natural language processing and artificial intelligence techniques that were aimed at solving problems related to intelligent information processing only to encounter more and more new obstacles along the way. The latest solution, the Semantic Web, appears as an open declaration of defeat: since natural language processing and AI techniques did not provide sufficient results, it is now proposed to put the burden on the shoulder of the authors of webpages who are expected to populate their pages with metadata and additional markup. Metadata is essentially a new form of controlled vocabulary; markup - at least in the form of XML, XSL, etc - is essentially a programming language. Existing studies of the use of controlled vocabularies and indexing practices in information science and studies of teaching programming languages to “everybody” (Python, 2002) have shown that both are difficult and full of unsolved problems. This can further dampen the expectations of the success of the Semantic Web.

In contrast to machines and despite numerous inter-cultural conflicts around the world, humans do communicate surprisingly successfully even across national, linguistic and cultural boundaries. The question then arises: why are humans successful at information processing tasks such as information integration, translation and communication, which computers find so difficult? One obvious answer is that human cognition is embodied and grounded in our shared experiences of living in the same world. AI researchers have theoretically explored the idea of symbol grounding in the early 1990’s but so far, connectionist artificial agents with perceptual interfaces have not been integrated with a large-scale capability of symbolic representations.
A critical review of existing strategies and theories of knowledge representation from a theoretical perspective is essential. The central question is: which strategies or theories have shown promise for which tasks, and why have they shown promise or have failed. It can be argued that successful knowledge representation requires a multitude of structures and techniques which each function with respect to specific conditions and contexts. This is in analogy to human cognition and communication, which also appears to be based on a multitude of structures. Some of these structures are "grounded" and thus provide for universals and primitives. Other structures are rule-based and provide for patterns, inferences and linguistic structures. Last but not least, meta-structures guide the combination and application of more basic structures. It can be argued that there is still hope that carefully designed systems that employ equally carefully designed knowledge representation strategies may overcome many of the current challenges for information systems. But this will require sophisticated theoretical models from a variety of disciplines.

2. The Semantic Web

In summer 2002 there were two workshops related to the Semantic Web co-located at Stanford University. The first one, the Semantic Web Working Symposium (SWWS, 2001) was funded by NSF, DARPA, INRIA and corporate sponsors; attracted several hundred participants from academia and industry; and had a highly technical focus. Many of the participants in this workshop were heavily involved in the implementation and development of the Semantic Web, such as invited speaker Eric Miller who is the W3C Semantic Web Activity Leader. The atmosphere at this workshop was very optimistic. The second workshop, the PORT Semantic Web Workshop (PORT-ICCS, 2001) attracted a much smaller number of participants but among them were famous researchers from diverse disciplines, such as Doug Engelbart, Ted Nelson, Terry Winograd, Keith Devlin, Tim Lenoir and Geoffrey Nunberg. The goal of PORT-ICCS was to investigate the use of the Semantic Web as a collaborative infrastructure for the development of Peirce Online Resource Testbeds (PORT). An unexpected outcome of this workshop was a fairly pessimistic view of the possibilities for the Semantic Web.

The SWWS final report defines the Semantic Web as "a vision: the idea of having data on the Web defined and linked in a way that it can be used by machines not just for display purposes, but for automation, integration and reuse of data across various applications" (SWWS Final Report, 2001). Eric Miller in his invited talk and introduction to SWWS stated that in the Semantic Web "everything has an online representation: people, places, things ...". According to the SWWS final report, the Semantic Web can be built: "in order to make this vision a reality for the Web, supporting standards, technologies and policies must be designed to enable machines to make more sense of the Web, with the result of making the Web more useful for humans." The SWWS participants thus see the Semantic Web as universal and technically feasible.

The more pessimistic views of the PORT-ICCS participants included the following: Terry Winograd remarked that only a small portion of the Web actually requires structure. Major parts of the Web have mostly recreational and entertaining use and probably do not require a Semantic Web. He further had doubts about the feasibility of mark-up; whether authors of web pages would be sufficiently
motivated to generate high quality, consistent mark-up. Ted Nelson characterized the Semantic Web as "just another delusion of the techies" because he considers human thought to be highly interconnected, non-hierarchical and difficult to process by machines.

In general there was suspicion among the PORT-ICCS participants that the Semantic Web really is a declaration of failure of the outcomes of artificial intelligence (AI) and natural language processing (NLP) research. If AI and NLP facilitated the creation of autonomous artificial agents with adequate language processing skills, then these agents would be able to process any information on the Web without requiring further mark-up. But neither AI nor NLP have yet reached their initial goals. For example, Geoffrey Nunberg, a linguist from Xerox Parc, argued at PORT-ICCS that although NLP can solve some tasks, such as genre classification and topic segmentation, universal NLP for any types of texts and unlimited domains will never work. The Semantic Web attempts to counteract these shortcomings by incorporating immense amounts of manual labor because it requires page authors to provide meta-information and mark-up. The Semantic Web is also under time pressure because a developing e-commerce industry demands instant solutions for conducting efficient, secure and profitable business on the WWW.

The question is what can and what cannot realistically be expected from the Semantic Web? What can the Semantic Web learn from successes and failures in the AI and NLP research of the last decades? Is the general idea of the Semantic Web pointing in the right direction, or are there better alternatives?

3. Successes and Failures

It is of interest to analyze which aspects of AI and NLP have produced successes and which have not so far. The success stories can be grouped into the following categories:

Pattern recognition: AI software has been built that can recognize patterns in visual, auditive, other sensory or abstract information. This ranges from face recognition to detecting credit card fraud (Kahn, 2002) by analyzing abstract patterns of credit card use. Pattern recognition also plays a part in speech recognition and robots (such as the Japanese toy dog robot that came on the market last year) because robots use pattern recognition as part of their perceptive interfaces.

Complex dynamic systems: AI software has been successful in modeling dynamic systems and designing solutions for controlling variables in such systems. For example AI solutions have been applied to monitoring tire pressure in cars, scheduling airport traffic, adjusting the movement in hand-held cameras, and building two-legged robots that walk in a human like fashion. These solutions often employ distributed representations. In a sense, even the search engine Google falls into this category. Although Google does not explicitly employ AI techniques, its success builds on exploiting the distributed nature of the dynamic linkage structure of the WWW.

Simulation of complex behavior: AI has achieved simulations of complex behavior such as language and human behavior. But these cannot yet pass the Turing test, i.e. simulate behavior that is indistinguishable from real human behavior. Examples are artificial agents that users of computer games engage with.
(Johnson, 2002) or "bots", such as the Alicebot, that can participate in on-line chat with human users.

Rule- and logic-based systems: these are AI systems, such as expert systems that encode facts in a formal representation (knowledge base or AI ontology) and compute inferences based on logical rules. Examples are chess computers and medical expert systems that support doctors in the diagnostic process (Kahn, 2002).

NLP techniques: NLP has been successful in achieving tasks that rely on linguistic pattern recognition and text summarization and classification. These are especially successful if applied to domains, which employ formalized language styles and vocabularies, such as medicine. Linguistic pattern extraction can be used to identify the language in which a document is written, for speech synthesis and to filter information. For example, Monster.com uses pattern extraction to filter job announcements and job applicants's information out of general WWW documents (Kahn, 2002). Data mining websites, such as KD nuggets (2002), list numerous software tools for clustering or classifying documents and similar tasks.

Failures: despite the achievements of NLP there are still no reliable tools available for translating between languages or for identifying the content of a document in more detail than a summary or classification. Less formalized linguistic domains, such as poetry and colloquial language, pose even more significant problems for any NLP tool. The Semantic Web attempts to remedy these failures by increasing the amount of information that can automatically be extracted from webpages.

4. Associative and Formal Structures

There are two types of conceptual representations: associative and formal ones (Priss (2001). Associative concepts are grounded (or embodied); they are dynamic and complex. Formal concepts are rule-based, abstract and aim for consistency. Human reasoning most likely involves both associative and formal structures (Sloman, 1996). But most of the AI and NLP techniques mentioned in the previous section are either exclusively formal (such as most NLP and rule- and logic-based AI) or exclusively associative (such as AI tools for pattern recognition, complex systems and simulations and statistical NLP tools). Maybe the failure of AI and NLP tools is due to the fact that they do not combine associative and formal structures? Humans shift seamlessly between both forms of representation. Clark (1997) argues that the combination of different representations can invoke feedback loops. Maybe these feedback loops are a defining factor in human cognition. Lakoff & Johnson (1999) argue that classical, formal philosophy can be misleading because it ignores the embodied nature of cognition. There can be no doubt that logical formalisms serve some purpose because, for example, most computer software is entirely logical-formal. Thus Lakoff & Johnson's argument for embodied structures is really an argument for the need to combine embodied, associative structures with formal ones.

It is interesting to observe that the Semantic Web in its current vision is entirely formal. Although it is intended to exploit the Semantic Web in data mining applications, many of which are associative, there is no discussion about employing associative structures directly in the creation process of the Semantic Web. For example, there are so far no suggestions to learn from Google's successes, which are based on the exploitation of the dynamic linkage structure of the web, and to
incorporate dynamic structures into the generation of the Semantic Web itself. Instead the creation of the Semantic Web relies heavily on the development of formal standards which become longer and more complicated with each new version. But studies in the area of controlled vocabularies, and formal programming languages (Python, 2002) and the experiences with 15 years of manual construction of the formal ontology CYC (Lenat, 2000) have shown that it is difficult (or impossible) to construct and manage large-scale formal systems that are consistent and user-friendly.

5. Conclusion

Instead of creating yet another entirely formal structure as an attempt to cope with natural language complexity, it might be more promising if the Semantic Web creators would pay more attention to previous successes and failures among AI and NLP tools. Most existing tools are either exclusively formal or exclusively associative. These tools are successful for limited tasks which are also predominantly formal or associative, respectively. But natural language and human cognition are more complex and require a combination of formal and associative structures (and possibly a multitude of other structures). It can thus be predicted that the Semantic Web also needs to combine both formal and associative structures and in general utilize multi-strategy approaches to knowledge representation. The current emphasis on complicated formal standards may not be sufficient. The underlying collaboration among authors that are expected to understand and apply these formal standards to their webpages is associative in nature and needs to be explicitly incorporated, modeled and understood by the supporters of the Semantic Web.

Notas
1. The following quotes were excerpted from notes taken at the Workshop. No formal proceedings have been published for PORT-I2CCS.

References

Universality and Basic Level Concepts

Abstract: This paper examines whether a concept's hierarchical level affects the likelihood of its universality across schemes for knowledge representation and knowledge organization. Empirical data on equivalents are drawn from a bilingual thesaurus, a pair of biomedical vocabularies, and two ontologies. Conceptual equivalence across resources occurs significantly more often at the basic level than at subordinate or superordinate levels. Attempts to integrate knowledge representation or knowledge organization tools should concentrate on establishing equivalences at the basic level.

1. Rationale

The degree of success attainable in the integration of multiple knowledge representation systems or knowledge organization schemes is constrained by limitations on the universality of human conceptual systems. For example, human languages do not all lexicalize the same set of concepts; nor do they structure (quasi-)equivalent concepts in the same relational patterns (Riesthuis, 2001). As a consequence, even multilingual thesauri designed from the outset from the perspective of multiple languages may routinely include situations where corresponding terms are not truly equivalent (Hudon, 1997, 2001). Intuitively, where inexactness and partialness in equivalence mappings across knowledge representation schemes and knowledge organizations schemes exist, a more difficult retrieval scenario arises than where equivalence mappings reflect full and exact conceptual matches.

The question we address in this paper is whether a concept's hierarchical level affects the likelihood of its universality/full equivalence across schemes for knowledge representation and knowledge organization. Cognitive science research has shown that one particular hierarchical level—called the basic level—enjoys a privileged status (Brown, 1958; Rosch et al., 1976). Our underlying hypothesis is that concepts at the basic level (e.g., apple, shoe, chair) are more likely to match across knowledge representation schemes and knowledge organization schemes than concepts at the superordinate (e.g., fruit, footwear, furniture) or subordinate (e.g., Granny Smith, sneaker, recliner) levels. This hypothesis is consistent with ethnobiological data showing that folk classifications of flora are more likely to agree at the basic level than at superordinate or subordinate levels (Berlin, 1992). The study reported here, which is only preliminary in nature, investigates the
validity of the hypothesis that basic level concepts are more universal than either
more general or more specific concepts in three contexts: across languages (e.g.,
between corresponding terms of the two languages of a bilingual thesaurus in the
social sciences), across vocabularies (e.g., between terms of different medical
vocabularies mapping to the same concept within the Unified Medical Language
System® [UMLS]; see related work by Bodenreider & Bean, 2001), and across
ontologies (e.g., between most nearly equivalent nodes of ThoughtTreasure and
WordNet; see related work by Hovy, 2002).

2. Basic Level: Privileged Level within a Hierarchy

The world is filled with a tremendously large number of individual concrete
entities. When we refer to any specific one of those entities, we usually do so using
a label that names a class the entity is a member of. We seldom bother, for instance,
to refer to an automobile by its vehicle identification number (VIN), which would
enable us to name it uniquely, but, given a neutral context, typically refer to it as a
car. Of course, the class of cars is not the only class that could be used in referring
to a specific automobile; both more specific (e.g., sedan) and more general (e.g.,
vehicle) classes/names are also available.

The hierarchical level typically chosen when we use language to refer to
entities turns out not to be random. A variety of processes by which we interact
with objects in our world converge on this same level, which is therefore dubbed
the basic level (Lakoff, 1987, 46-47). Several of these converging processes are
linguistic in nature. In addition to the reference process just mentioned, names for
basic level categories tend to be shorter than are the names of superordinate and
subordinate categories. Words for basic level categories tend to enter the language
earlier and to be learned by children earlier than words naming more general and
more specific classes. Other processes that privilege this basic level of the hierarchy
concern perception, function, and knowledge organization. On the perceptual level,
the basic level is the highest level in a hierarchy where humans can normally form a
single mental image of a class and where they can identify the class by its average
shape. As to function, the basic level tends to be the highest level in a hierarchy
where humans interact with entities with a relatively constant motor program. As to
knowledge organization, when people are asked to list all that they know about a
certain category, the biggest increase in number of statements, over and above what
one can say about the next-most-general superordinate class, comes at the basic
level, thus implying that more information is stored that is specific to the basic level
than to any other hierarchical level.

The basic level concept arose in the context of classifying physical objects. It
is not always clear how to apply some of the converging processes noted above,
especially those related to perception and function, to abstract concepts, processes,
events, and so forth. We will assume, however, that the basic level concept does
apply to these less concrete contexts and that the several linguistic criteria that tend
to converge for concrete entities will likewise converge for non-concrete entities.
3. Equivalence across Knowledge Organization and Knowledge Representation Schemes

Our goal is to investigate whether the privileges of the basic level phenomenon extend to knowledge organization and knowledge representation schemes. We start by taking a random sample of (usually) a dozen terms or nodes from each of three types of schemes (a bilingual thesaurus, a metathesaurus to which a number of vocabularies have been mapped, and a set of ontologies), expanding each term to include all hierarchically related terms/nodes. For each term in the source scheme (including both the randomly selected terms/nodes and the terms/nodes identified through hierarchical expansion; each of the two tools being compared is the source scheme for half of the investigation and the target scheme for the other half), we identify the term/node in the corresponding target scheme that is the closest equivalent. We then analyze these mappings for degree of equivalence, taking into account the semantic scope of the terms at each terminus of the mapping, as well as the hierarchical placement of these terms within their home vocabulary/ontology. We also use available linguistic information (term length, date term entered language, etc.) to determine the basic level within each hierarchy. A goodness-of-fit test is then applied to the results of the analysis to determine if there is a significant difference between the degree of equivalence in mappings—that is, the universality—of subject terms at basic and non-basic levels. In this analysis, the degree of equivalence between a concept in the source scheme and its nearest concept in the target scheme is characterized according to five categories: Equivalent or nearly equivalent, More general/specific, Semantic type mismatch, Not a good match, and Missing (from target).

3.1. Equivalence across Languages: Canadian Literacy Thesaurus/Thésaurus canadien d’alphabétisation Case Study

The Canadian Literacy Thesaurus / Thésaurus canadien d’alphabétisation (CLT/TCA, 1996) is a bilingual list of subject terms (in French and English) relating to the field of adult literacy. The French portion of the thesaurus includes 1950 descriptors, the English 1890. The two lists of terms were developed and structured independently, then later reconciled. Thus, we would expect to find a large percentage of exact equivalences across matching terms on the two sides of the thesaurus. However, despite the reconciliation, not every term in one language has an equivalent term in the other language. Nor are “equivalent” terms always fully equivalent.

Because of the degree of specialization represented by the thesaurus, the hierarchies tend to be shallow, often only two levels deep. Where this is true, it is clearly not the case that subordinate, basic-level, and superordinate terms are all present. Indeed, it is not even always the case that basic-level terms are present. Because of this data sparsity and the relatively small size of the vocabularies involved, a random sample comprising 1% of the descriptors from each language was drawn, rather than randomly selecting six descriptors from each vocabulary. Also, presumably because of the reconciliation process that the thesaurus underwent, some number of the “exact matches” that occur are fabricated: The descriptor in one language is generated on the basis of a word-for-word translation of the descriptor from the other language, although the phrase is not attested in standard terminological resources. Such exact matches have been categorized as forced equivalences and are conceptually closer to missing terms than to equivalent
terms. The chart below summarizes degree of equivalence for unique subordinate, basic level, and superordinate terms as compared with their most nearly equivalent concepts across the language divide of the Canadian Literacy Thesaurus / Thésaurus canadien d’alphabétisation.

<table>
<thead>
<tr>
<th>Degree of equivalence</th>
<th>Subordinate-level concepts</th>
<th>Basic-level concepts</th>
<th>Superordinate-level concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equivalent (or nearly so)</td>
<td>18</td>
<td>22</td>
<td>20</td>
</tr>
<tr>
<td>More general/specific</td>
<td>7</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Semantic type mismatch</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Not a good match</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Forced equivalence</td>
<td>6</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Missing from target</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

3.2. Equivalence across Terminologies: Unified Medical Language System Case Study

The Metathesaurus® of the Unified Medical Language System (UMLS) is a tool that, in its current (2002AA) version, links 2.1 million terms from over 60 biomedical vocabularies to a set of 776,940 concepts. The mapping of each individual vocabulary to these concepts generates, in effect, mappings between any two of the included vocabularies. This study is based on mappings between anatomical terms from two widely-used comprehensive multiaxial hierarchical medical vocabularies: MeSH ® (2002) and SNOMED International ® (1998). MeSH (Medical Subject Headings) is maintained by the U.S. National Library of Medicine, and the 2002 version contains over 122,000 main headings (19,000) and supplementary concept records (103,500). SNOMED (Systematized Nomenclature of Medicine) is maintained by the College of American Pathologists; SNOMED International (1998) contains over 156,000 concepts.

The six hierarchies selected from MeSH as source have as basic level terms tooth, uterus, arm/hand/finger, gland, pancreas, and blood vessel; the six selected from SNOMED as source have as basic level terms abdomen, eye, bone, nose, cavity, and bone marrow. It should be noted that hierarchies within anatomy are often based on the part-whole relationship rather than on the subsumption (taxonomic) relationship. Consequently, the basic level terms just indicated, which are based on taxonomic hierarchies, do not always appear in the hierarchies given in MeSH and SNOMED, but have been identified using WordNet (see section 3.3). The chart below summarizes degree of equivalence for unique subordinate, basic level, and superordinate concepts as compared to the cross-thesaural most nearly equivalent concepts for MeSH and SNOMED.
3.3. Equivalence across Ontologies: ThoughtTreasure and WordNet Case Study

ThoughtTreasure (http://www.signiform.com/tt/htm/tt.htm) describes itself as “a comprehensive platform for natural language processing... and commonsense reasoning.” Its knowledge base includes a lexicon of 25,000 hierarchically organized concepts, to which 55,000 English and French words have been mapped. WordNet (http://www.cogsci.princeton.edu/~wn) is an English language lexical database built around sets of synonymous word senses (synsets), each of which corresponds to a node in a tree structure. WordNet organizes 190,000 word senses into 110,000 synsets; separate tree structures exist for nouns, verbs, adjectives, and adverbs. The ontologies of such other knowledge representation systems as Cyc, Sensus, and Ontosaurus are admittedly more typical of knowledge representation systems, but the linguistic basis of both ThoughtTreasure and WordNet permit the identification of basic level within their hierarchies using the lexical criteria outlined in section 2, which otherwise would not be possible. They also have the very practical advantage of being freely available in their entirety.

The six randomly selected hierarchies based on ThoughtTreasure as source have as basic level terms trial, wife, phone number, column, resistor, and art; the six based on WordNet as source have as basic level terms blood vessel,2 word, training, official, building, and store. The following chart summarizes degree of equivalence for unique subordinate, basic level, and superordinate concepts as compared to the most nearly equivalent concepts identified across ThoughtTreasure and WordNet.

<table>
<thead>
<tr>
<th>Degree of equivalence</th>
<th>Subordinate-level concepts</th>
<th>Basic-level concepts</th>
<th>Superordinate-level concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equivalent (or nearly so)</td>
<td>6</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>More general/specific</td>
<td>2</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>Semantic type mismatch</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Not a good match</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Missing from target</td>
<td>7</td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>
4. Discussion and Analysis

The cumulative results from the three case are captured in the chart below (the Forced equivalence category from the first case study has been folded into the Missing from target category here):

<table>
<thead>
<tr>
<th>Degree of equivalence</th>
<th>Subordinate-level concepts</th>
<th>Basic-level concepts</th>
<th>Superordinate-level concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equivalent (or nearly so)</td>
<td>41</td>
<td>42</td>
<td>39</td>
</tr>
<tr>
<td>More general/specific</td>
<td>12</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>Semantic type mismatch</td>
<td>2</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Not a good match</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Missing from target</td>
<td>22</td>
<td>0</td>
<td>14</td>
</tr>
</tbody>
</table>

Statistical ($\chi^2$/goodness-of-fit) tests were applied to these results to investigate the basic hypothesis that conceptual equivalence is more likely to occur in knowledge organization and representation tools at the basic level than at other levels. The hypothesis is specifically confirmed as follows:

1. Across the three case studies combined, basic level terms are more likely to have equivalents in other knowledge organization and representation tools than are non-basic level terms (subordinate and superordinate terms combined) ($\chi^2 = 25.24047$, df = 1, $\alpha = .001$).

2. Across the three case studies combined, basic level terms are more likely to have equivalents in other knowledge organization and representation tools than are subordinate terms ($\chi^2 = 21.45006$, df = 1, $\alpha = .001$).

3. Across the three case studies combined, basic level terms are more likely to have equivalents in other knowledge organization and representation tools than are superordinate terms ($\chi^2 = 22.41596$, df = 1, $\alpha = .001$).

5. Conclusion

The confirmation of our hypothesis that full equivalence between tools occurs more often at the basic level than at either superordinate or subordinate levels for both knowledge representation schemes and knowledge organization schemes should illuminate future efforts to build universal tools of subject or semantic access. Specifically, it suggests that crosswalks between such tools should focus on mappings at the basic level, without attempting to impose a comprehensive mapping at all hierarchical levels.

Notes
1. The use of a part hierarchy results in three basic level terms appearing there.
2. In this case the lexical criteria are mixed, some pointing to vessel as the basic level and others to blood vessel.
References


Grant Campbell
Faculty of Information and Media Studies, University of Western
Ontario, London, Ontario, Canada

Chronotope and Classification: How Space-Time Configurations Affect the Gathering of Industrial Statistical Data

Abstract: Bakhtin’s theory of the chronotope is used to examine how representations of space and time affect the first six classes of the North American Industrial Classification System. By examining the class sequence as a narrative of the product life cycle, the study suggests that this new classification system, designed to harmonize the gathering of statistical data among the three countries of North America, manifests an economic paradigm which diminishes the visibility of community ties based on geographical proximity, community identity, and communication across social and economic barriers.

1. Introduction

This paper was inspired by a challenge and a metaphor, both of which appeared in Bowker and Star’s recent work on classification (2000). First, the challenge. Human society suffers, the authors argue, from “an impoverished vocabulary for collective moral passages” (Bowker and Star, 2000, p. 6), and this prevents us from completely understanding classifications, which, in their complex evolution and maintenance, require us to comprehend moral activity that works by long-term accretion and collaboration, rather than by an individual’s rational choice. Second, the metaphor Bowker and Star frequently refer to classification systems as stories: “standard narratives that appear universal” (Bowker and Star, 2000, p. 41), but which are really a series of choices, each of which silences one voice and valorizes another.

This paper attempts to use the metaphor to rise to the challenge. By treating classifications specifically as narratives, and subjecting them to narrative theory as developed in the humanistic disciplines, classification theorists could enrich the vocabulary of collective moral passages, and provide the tools for more sophisticated and telling critiques, not just of classification systems, but of the social, political and economic power structures that create and sustain them. What follows is a test example, in which I analyze an excerpt from the North American Industrial Classification System, and explore how Mikhail Bakhtin’s narrative theory of the chronotope enables us to uncover assumptions and patterns that might otherwise be overlooked.

2. The Social Context of Classification

Modern classification theory readily acknowledges that classifications are social constructs. Current theorists have analyzed classifications as artifacts embedded in particular information cultures (Hjørland and Atbrechtsen, 1999), and
as products of a social context, rather than an exclusively cognitive context (Jacob, 131). This social context gives the classification its intellectual coherence, and constrains the kinds of questions that can be posed within it.

However, these assumptions can be difficult to trace and acknowledge, partly because they manifest themselves through complex networks of rules and guidelines. Second, these assumptions are frequently expressed, not through what is spoken or written, but by subtle interactions between speech and silence, between explicit and implicit, similar to the interaction between speech and silence which, according to Foucault, characterizes Western society’s treatment of sexuality: the “perpetual spirals of power and pleasure” (Bowker and Star, 2000, p. 45).

Some of these assumptions may emerge if we treat a classification system as a narrative: a collection of incidents in a particular sequence. This is hardly an unusual approach. Narrative theory has been used successfully by various scholars examining genre in electronic documents (Smoliar and Baker, 1997), text linguistics (Beghtol, 1986) and discourse analysis (Beghtol, 1997). Furthermore, by treating a classification as a narrative sequence, we can also use related classification theories of sequence: in particular, Ranganathan’s principle of “Helpful Order” in an array.

3. Classification Theory: Ranganathan and the Concept of “Helpful Order”

The most obvious correlation between narrative and classification theory lies in Ranganathan’s principle of class sequence. “The sequence of the classes in any array,” he argues, “should be helpful” (Ranganathan, 30). In the various principles he enumerates for such sequences—decreasing extension, increasing concreteness, spatial contiguity, canonical sequence—he assumes that the user moves through a systematic arrangement in linear fashion, and that that linear progress should make some kind of sense, even if it is only dimly perceived.

Time and space play prominent roles in this principle of helpful sequence, in two ways. First, by assuming that the user moves along the shelves in sequence, Ranganathan assumes that helpful order involves determining whether a user encounters a subject sooner or later in the browsing activity. Furthermore, the sequence ensures that certain facets will be encountered next to each other on the shelves, while others will be spatially separated.

Second, both time and space are important determinants of order in themselves: principles such as evolutionary order and spatial contiguity assume that a meaningful arrangement often involves placing documents in chronological sequence, or by geographical location. A topic such as “India during the time of British Rule,” for instance, would have two temporal dimensions to it: the period in India’s history that corresponds to British Rule, and the moment at which the user would encounter such a topic in a horizontal movement through a sequence of documents. Similarly, a topic such as “education in Great Britain” would refer both to the geographical area known as Great Britain, in relation to other geographical areas, and the specific location on the shelves in which that topic may be found, in relation to documents on education in other areas.
4. Narrative Theory: Mikhail Bakhtin and the Concept of the “Chronotope”

Since time and space are prominent structuring principles in helpful order, a narrative theory that specifically addresses space-time configurations would be useful. Mikhail Bakhtin offers such a specific concept with the “chronotope”: “the intrinsic connectedness of temporal and spatial relationships that are artistically expressed in literature” (84). In his discussion of ancient narrative genres, Bakhtin isolates three distinct chronotopes which correspond with three primary types of story:

- The adventure novel of ordeal, in which the protagonist leaves the normal world to move through a dream-like landscape, experiences a variety of random adventures, and then returns to the everyday world at the same moment he left;
- The adventure novel of everyday life, in which the hero moves through the everyday landscape in chronological sequence, but who experiences events as a series of disruptions and transformations in the ordinary nature of things;
- The biographical narrative, in which the protagonist again moves through the everyday landscape in chronological sequence, but who experiences events as a process of continuity and development.

If we transfer this paradigm to the context of classifications, and particularly to the order across classes of the system, we can pose 4 primary questions:

- In the helpful sequence, who or what is the “protagonist” of the development across the classes?
- Is there a temporal principle in the order of classes, which assumes that time is passing between one class and the next?
- Is the sequence that governs the movement from one class to the next based on continuity or disruption?
- Does the order of the classes relate to the “everyday world,” or does it set up an abstract space?

5. The North American Industrial Classification Scheme

5.1 Background

The North American Free Trade Agreement (NAFTA) was a highly controversial arrangement to lower trade barriers between Canada, the United States and Mexico. The deal was signed in 1994, over the vociferous opposition of such groups as the Council of Canadians, who argued that Canada, by entering into the agreement, was losing its economic and cultural sovereignty.

As part of NAFTA, the statistical agencies of the United States, Canada and Mexico have agreed to adopt a harmonized classification system for the collection of statistical data. This replaces Canada’s 1980 Standard Industrial Classification, and was designed with two primary objectives. First, although industrial classifications are used for a wide variety of purposes, ranging from legislation to employment searches, the primary purpose of NAICS is to gather statistics: “Government departments and agencies and other users that use it for
administrative, legislative and other non-statistical purposes, are responsible for interpreting the classification for the purpose or purposes for which they use it" (Statistics Canada, 1999). Second, it is designed to harmonise the data between the three members of NAFTA, so that all three countries are collecting economic data in a way that facilitates comparison across the three member countries.

If we compare the 1980 Standard Industrial Code with the North American Industrial Classification according to the four questions posed by Bakhtin’s chronotope theory, some intriguing differences come to light.

5.2 Who or What is the Protagonist?

According to the NAICS documentation, previous classifications, such as the Canadian SIC 1980, “have all used mixed criteria” when defining industries in their classification (Statistics Canada, 1999). One of the achievements of NAICS is the clarification of the main unit of classification as the “producing unit”:

NAICS is based on a single production-oriented concept. Producing units are grouped into industries according to similarities in their production processes. The boundaries between industries demarcate, in principle, differences in production processes and production technologies. This means that, in the language of economics, producing units within an industry have similar production functions that differ from those of producing units in other industries. (Statistics Canada, 1999)

If, however, we examine the sequence of classes in the first half of the NAICS classification, we see a different narrative and a different protagonist. NAICS divides roughly into two halves, the first devoted to production of goods, and the second to provision of services. The production of goods has 6 main classes as follows:

11 Agriculture, Forestry, Fishing and Hunting
21 Mining and Oil and Gas
31 Manufacturing
41 Wholesale
44-45 Retail
48-49 Transportation and Warehousing

In terms of the class sequence, the protagonist is not the “producing unit,” but rather the economic commodity, working its way from primary production, by cultivation or extraction, through the manufacturing process to its distribution and sale. This is a familiar pattern, similar to the classification of the Book Sciences in the Library of Congress Classification. But it suggests that the explicit unit of measure in the classification documentation is not the same as the unit of narrative sequence, at least in the first half of the classification.

5.3 Temporal Context

Because the classification follows an evolutionary sequence in the first half of the schedules, the temporal context is roughly continuous for the product side of the classification. The sense of temporal continuity has been enhanced with NAICS by the collapse of several primary industries into one category that had previously held separate ones. In SIC 80, the first four classes were assigned to agriculture, fishing and trapping, logging and forestry, and mining respectively. In NAICS, these have been collapsed into two categories, thereby launching the temporal movement from extraction to refinement more quickly and obviously.
5.4. Sequence

Despite this smooth progression, however, the NAICS paradigm undergoes an abrupt transformation from production to services; beginning with the information and cultural industries, the classification shifts into service industries such as finance, real estate, education, health care, arts and entertainment and other services, with Public Administration bringing up the rear.

5.5. Spatial Context

By defining the producing unit as the unit of measure, NAICS has introduced subtle changes to the spatial dimensions of industry statistics as enshrined in SIC 80. Because producing units are grouped according to similarities in production processes and production technologies, certain activities have been displaced from their previous contexts. In some cases, the industry is moved to another part of the product's narrative sequence. "Bait preparation and supply service," for instance, which was formerly classed together with fishing in the Fishing and Trapping Industries, has been displaced to the Wholesale Trade at class 41. Similarly, "Production of transportable goods," formerly part of Mining, has been moved to Manufacturing. In such cases, the industry has been relocated to another stage of the narrative sequence, but has not been displaced entirely.

In other cases, the industry has been moved to the other side of the product/services divide, and has been expunged from the production narrative entirely. In Agriculture, for instance, veterinary services, soil and seed testing services, agricultural consultants and agricultural research and development have been displaced to Class 54: Professional, Scientific and Technical Services. Fisheries research and development has been also been moved from Fishing to Category 54; the repair of fishing gear now lives in Class 81: Other Services, and Fishery Patrols are now in Class 91: Public Administration.

This separation of product and service creates a rupture in NAICS between the spatial dimensions of arrangement within the classification and the spatial dimensions of the physical, social and cultural world in which these industries exist. By removing services into their own area, NAICS provides a highly abstract picture of economic relationships as social and cultural phenomena. By depicting agriculture as a series of activities divorced from veterinary services and agricultural research, and fishing as an activity distinct from fisheries research, NAICS-classified statistics present an economic picture which radically diminishes the visibility of community ties based on geographical proximity, community identity, and communication across economic barriers.

6. Conclusion

It would be absurd to equate classification systems and literary narratives, and insist on a one-to-one correspondence between an ancient Greek romance and a twenty-first-century industrial classification system. Nonetheless, Bakhtin's theory of the chronotope does, I believe, provide us with some tentative tools for enhancing our vocabulary of collective moral activity. By analyzing the narrative structures inherent in the class order of NAICS, we may provide those who are urgently querying globalization trends with fresh means of exploring and articulating their reservations and concerns.
References
Marianne Lykke Nielsen  
Royal School of Library and Information Science, Copenhagen, Denmark  

Anna Gjerluf Eslau  
H. Lundbeck, Valby, Denmark  

Corporate Thesauri – How to Ensure Integration of Knowledge and Reflection of Diversity  

Abstract: This paper evaluates and compares three thesaurus construction methodologies: literary scanning, word association tests, and involvement of subject expert groups. The evaluation concentrates on exploring advantages in relation to the sub-processes: collection, formation and structuring of concepts and terms. Quantitative as well as qualitative analyses have been carried out. The analysis shows that the methods are complementary each providing distinct conceptual information from respectively a domain-oriented and a scientific viewpoint. The combination of methods provides a thesaurus, at the same time, mapping authoritative language and reflecting the diversity of language.  

1. Introduction  
Basically, a corporate thesaurus serves as a mechanism to facilitate communication, improve learning and support information retrieval across the organisation. Recently, in relation to knowledge management and implementation of corporate intranets, portals and document management systems, there has been an increased interest in conceptual tools like thesauri providing “... a semantic road map to individual fields and the relationships among fields, to relate concepts to terms, and to provide definitions ...” (Soergel, 1996. p.167). Business corporations consist of divisions and subsidiaries situated in various locations, nationally and internationally, with staff coming from different knowledge domains, having different cultural, social, educational and professional backgrounds. In large organisations there exists a range of discourse communities, each of which having specific tasks and objectives, generating their own jargons (Cabre, 1999; Morgan, 2000, Gilchrist and Kibby, 2000).  

In corporate information environments, conceptual tools like thesauri should describe the languages in play and bridge the boundaries between the individual user and the set of jargons. The compiler of such tools faces and must consider ambiguity problems like subject crossover and overlap, interdisciplinary borrowings, disciplinary or domain-specific focus, cultural differences, linguistic differences demanding thorough knowledge about the corporate language use (Gilman, 1997).  

The present paper compares thesaurus construction methods, based on literary warrant, with methods based on user warrant with the aim to develop a framework for thesaurus construction and maintenance combining the two approaches. The paper is structured as follows: Section 2 discusses the conditions
and problems related to thesaurus construction. Section 3 describes the purpose of the research project and the evaluation method, and Section 4 discusses the empirical findings. The concluding Section 5 outlines a framework for thesaurus construction combining literary- and user-oriented construction methods.

2. Thesaurus construction

Thesaurus construction involves three primary processes: collection, formation and structuring of concepts and terms. Principles for each sub-process have been developed and embodied in a variety of standards and guidelines (see e.g.: Soergel, 1974; International organization for Standardization, 1986; Aitchison, Gilchrist & Bawden, 2000). Particular problems are related to each process, and the guidelines suggest alternative solutions. However, it is often domain-related factors, which make it possible to choose among the alternatives. This implies a thorough knowledge of the information environment and the discourses and languages in play.

Professional jargons are specialized languages, based on and derived from general language, presupposing special education and restricted to communication among specialists in the same closely related fields (Cabre, 1999). Professional jargon is characterised of being well defined with little ambiguity of meaning and use. Ideally, each term of the vocabulary refers to one concept, and concepts are named by only one term. Relationships among concepts are clear and well defined. However, within the same organisation several jargons will be at work, and it may be difficult to capture, separate and describe the domain-specific languages in play and present them in an understandable form. Also “double-coded terms”, terms adapted from everyday vocabulary to professional, specialized languages with a changed meaning, may cause problems as corporations consist of professionals approaching the subject domain with different levels of perspective and subject knowledge (Haas and Hert, 2000).

The thesaurus compiler face several problems. Concerning collection it is difficult to define the topical focus and decide what concepts and terms to include. Decisions regarding form pose other problems: grammatical forms, choice of preferred terms, and concept definitions. Concerning structuring it is difficult to capture perspectives of the organisation and relate concepts and terms so that the relationships provide conceptual understanding of use, mapping subtle, but important distinctions and relationships. Many concepts are poly-hierarchical, and deciding what viewpoints to reflect by the hierarchies is difficult avoiding both information overload and mislead. The level of specificity of hierarchies constitutes another important problem.

Sources for collection of conceptual information for thesaurus construction are normally divided in written, literary sources and unwritten sources, for instance, knowledge obtained from experts and users by interviews, workshops, etc. Each methodology provides distinct information, supports the thesaurus compiler differently, and involves different competences and resources processing the information collected. Standardised sources require less effort in gathering and collecting well-defined disciplinary-oriented and authoritative information. Unwritten, non-standardized sources make possible collection of a set of terms reflecting the information environment actual usage from a more current viewpoint (Soergel, 1974).
3. Purpose and evaluation method

The present research project aims at evaluating methods respectively based on literary warrant and on user warrant. The hypothesis behind the project is that nature of conceptual information gathered by literary sources and human sources differ as regards topical perspective, specificity and resources for processing. The project focuses on investigating the nature of data collected in relation to the three primary processes in thesaurus construction; collection, formation, and structuring. The information collected is compared and strength and problems of the two (literary and human) sources to conceptual information is evaluated.

The analysis includes the following methods: a) scanning of internal or external literary sources; b) word association tests, and c) cooperation with groups of subject experts (Soergel, 1974, Aitchison, Gilchrist and Bawden, 2000, Lykke Nielsen, 1997).

The three methodologies under investigation have been used for thesaurus construction in a real-life thesaurus design project in a large research and development company within the pharmaceutical industry. The methods have been applied and evaluated according to findings of a domain study exploring purpose and role of the corporate thesaurus. According to the domain study the objective of the thesaurus is twofold. The thesaurus shall support information retrieval by referring to retrieval synonyms (strongly related terms appropriate for query expansion), and the thesaurus shall support learning and knowledge exchange providing insight into the specific jargons used within the corporation. The thesaurus should mediate the preferred corporate language and the internationally recognized scientific and authoritative language, but the thesaurus should also expose the diversity and vocabulary developments over time (Lykke Nielsen, 2001).

The literary scanning aimed at identifying and comparing vocabularies recognised within the corporation and internationally. Primarily, glossary lists were revised. Concepts were picked in order to cover the standardised vocabulary, and preferred terms and definitions were chosen to reflect the scientific, authoritative language use. The first version of the corporate thesaurus was constructed on basis of the literary scanning.

The purpose of the word association tests was to gather retrieval synonyms and capture term relationships reflecting the viewpoint of the work domain. A primed, continued test methodology was used in order to capture as many relationships as possible. The respondents, representing four information-oriented work task groups within the area of research and development were asked (primed) to make associations in relation to their work tasks and frame of interest. A total of 100 descriptors from the literary-based thesaurus were presented to the thirty-five respondents, providing a total of 2885 associative responses. Twenty stimuli words, of high relevance for the work domain, were presented to the full set of respondents, the rest of the stimuli were only presented to a single work task group.

Four expert groups representing the main business areas were established to support the thesaurus manager in identifying the relevant topical areas, in choosing preferred terms and evaluating the structure as regards specificity and relationships. The members were selected according to work domain experience and decision competence concerning vocabulary control. Formal group meetings were held regularly, and experts were consulted ad hoc by e-mail or phone.
Sub-processes of thesaurus construction:

<table>
<thead>
<tr>
<th>Sub-processes of thesaurus construction:</th>
<th>Issues related to thesaurus construction:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collection</td>
<td>Topical coverage</td>
</tr>
<tr>
<td>Domain orientation</td>
<td></td>
</tr>
<tr>
<td>Formation</td>
<td>Grammatical forms</td>
</tr>
<tr>
<td>Selection of preferred terms</td>
<td></td>
</tr>
<tr>
<td>Definition</td>
<td></td>
</tr>
<tr>
<td>Structuring</td>
<td>Type of relationships</td>
</tr>
<tr>
<td>Specificity</td>
<td></td>
</tr>
<tr>
<td>Synonym variations</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1. Evaluation criteria

The evaluation was carried out in three steps. First a quantitative comparison was made measuring the overlap as regards relationships between the literary-based thesaurus records and the 100 records based on the word association tests. The aim of this analysis was to evaluate whether relationships identified by the two methods differ. The second evaluation was qualitative, analysing the nature of the terms and relationships discovered by the word association tests, hereby investigating the differences between the literary scanning and the word association method. In the third step the conceptual information provided by the expert groups was compared to the other methods. The comparison was founded on the evaluation criteria presented in Table 3.1.

4. Discussion of empirical findings

Concerning *collection* the literary scanning of external sources provided a broad, general set of terms requiring subsequent selection of domain-relevant concepts. The scanning of internal documentation provided a more domain-oriented subset of concepts. 54% of the literary-based term relationships established in the first version of corporate thesaurus were also elicited by the word association tests (see Table 4.2), indicating that the two methodologies collect a high amount of similar conceptual information. However, the word association method elicited more up-to-date conceptual information reflecting the latest research and present focus of the corporation. Many relevant concepts were provided by the word association tests, confirmed by the fact that 92% of the new, unique terms (397 terms), discovered by the word association method, were later included in the corporate thesaurus. Neither of the methods mentioned provided a complete, exhaustive picture, which became clear when expert groups were revising the thesaurus. Expert groups have an important role controlling whether the thesaurus covers the viewpoint of the corporation. The word association method also provided a considerable amount of useless information. In the present project 31% of the word associations to the 20 stimuli qualitatively evaluated, were considered useless: word associations not possible to verify, broad concepts out of context, too specific concepts, synonym variations in other languages or curious translations, and phrases. In a professional environment this kind of terms is not useful as retrieval synonyms.
Concerning formation, the word association tests as well as the expert group evaluation revealed a pattern of grammatical forms for individual concepts, but it was not possible to identify a general pattern, and thesaurus construction guidelines have been used for the corporate thesaurus. As regards choice of preferred terms, the literary-based thesaurus followed international vocabularies selected by the expert groups. However, the later expert group evaluation revealed that the literary sources were sometimes obsolete as regards current naming. An analysis of word associations showed that the frequent word associations (associated by three or more respondents) tend to follow the preferred terms found by literary warrant, on average 58%, indicating that choice of preferred terms might be based on word associations, especially because the word associations furthermore reflected up-to-date language use. Definitions are an important part of a corporate thesaurus supporting learning and knowledge exchange. In a highly regulated environment like the present, it is natural to base definitions on authoritative and scientific literary sources. However, the word association tests revealed domain-specific knowledge, making it possible to work up more domain-oriented definitions compared to the generic, "neutral" definitions found in the literature. From the word associations, for example, given to the stimuli interactions, the thesaurus manager became aware of the connection between interactions and cytochrome P-450. The word associations inspired her to investigate further the relationship between interactions and cytochrome P-450. The investigation had as result that the concept cytochrome p-450 has been included in the corporate thesaurus, and that the definitions have been revised according to the findings (see Table 4.1).

Concerning structuring the hierarchical relations was best found by the literary scanning. International sources, like the MESH thesaurus, provided exhaustive and well-defined hierarchies for many of the concepts covered by the corporate thesaurus. For other concepts no appropriate sources existed, and therefore expert groups played an important role in the hierarchical structuring. Some domain-specific concepts, e.g., corporate drug products, require hierarchies tailored to the context and may only be developed by domain experts. Concerning degree of specificity of hierarchies, the preliminary domain study identified the most important concept categories to be represented with a high degree of specificity. The expert groups also provided useful information about specificity. Neither the word association tests nor the literary scanning provided such information. The literary sources did not reveal a high amount of associative relations, whereas the majority of the word associations consisted of related terms. The qualitative analysis showed that the total number of BT was 10%, the number of NT is 8%, the number of synonyms 5%, and the number of RT 46%. The relationships identified by the word association tests covered 54% of the relations established from the literature scanning (see Table 4.2 for coverage of literary-based relationships).
Drug interactions
The action of one drug on the metabolism, effectiveness or toxicity of another. (MESH)

Drug interactions
The action of one drug on the metabolism (see cytochrome P-450), effectiveness or toxicity of another (MESH red)

Cytochrome P-450
Cytochromes P450 are the principal enzymes for the oxidative metabolism of drugs and other xenobiotics. Among the xenobiotic-metabolizing cytochromes P450, five forms, CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4 appear to be most commonly responsible for the metabolism of drugs. Inhibition of cytochrome P450-mediated metabolism is often the mechanism for drug-drug interactions. The potential for enzyme inhibition is routinely assessed by performing in vitro inhibition studies using CDNA-expressed enzymes or human liver microsomes. These types of studies are becoming a routine part of the drug registration data package. (red)

Table 4.1. Revision of definitions according to findings of word association tests.

<table>
<thead>
<tr>
<th>Type of relationships</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>BT – Broader terms</td>
<td>68%</td>
</tr>
<tr>
<td>NT – Narrower terms</td>
<td>52%</td>
</tr>
<tr>
<td>RT – Related terms</td>
<td>52%</td>
</tr>
<tr>
<td>SYN – Synonyms</td>
<td>45%</td>
</tr>
<tr>
<td>ALL - All relationships</td>
<td>54%</td>
</tr>
</tbody>
</table>

Table 4.2. Literary-based relationships covered by the word associations for the 100 stimuli.

These findings show that a high amount of similar conceptual information is collected by the two methods compared. In addition, a large amount of unique associative relations not discovered by the literary scanning was elicited by the word association method. These relationships played an important role as mediator of up-to-date, domain-oriented knowledge. Both literary scanning and the word association method identified a number of equivalence relationships. The expert groups were used to revise the synonyms according to the domain-specific language use.

5. Conclusion and future work

The evaluation of the three thesaurus construction methods has clearly shown that they complement each other having different forces. Our work shows that a construction process should always be based on a domain study clarifying the
purpose and role of the thesaurus. The main concept categories will be identified by the domain study, and the collection process should be initiated by carrying out word association tests based on stimuli words covering the main concept categories. The stimuli words should preferably be selected from up-to-date documentation, strongly related to the topical scope of the thesaurus. Expert groups are recommended to assist the thesaurus compiler in selecting literary sources on which the subsequent evaluation of relevance and validity of the word associations should be based. One may say that the word associations act as “entrance” to the current corporate languages in play. The word association tests will not cover exhaustively the topics, and must be supplemented with conceptual information from literary scanning of internal and external sources.

All the methods evaluated require considerable human intellectual resources. The challenge is to combine the methods in such a way that the advantages of each method are exploited optimally. By joining the methods it is possible to construct a corporate thesaurus, which, at the same time, map the diversity of the corporate language use, integrate corporate knowledge and refer to relevant retrieval synonyms.

Another important issue to be addressed is thesaurus maintenance: how to ensure that the corporate thesaurus at all times exposes vocabulary developments. We believe that word association tests, due to their ability to elicit up-to-date, domain-oriented knowledge, performed on a regular basis will be the ideal tool in the future maintenance of the corporate thesaurus.

Future research will concentrate on developing the word association methodology and create word association tests in electronic form to be sent out by e-mail. The research will furthermore focus on investigating the possibilities of collecting conceptual information other than word associations.

References


Abstract: A work in progress. An analysis of the handling of the domain “environmental science” in three universal classification schemes. Attention is given to problems of primary location, interdisciplinarity, degree of scatter, terminology and structure of the domain.

1. Introduction
Many traditional tools for organizing and providing access to recorded knowledge have become increasingly inadequate in responding to the changing requirements for effective organization and retrieval. The universal classification systems have been primary targets for criticism. In varying degrees they have been proven too rigid and somewhat limited as tools of modern knowledge organization. These inadequacies can be attributed to several significant factors - the changing nature of knowledge itself, the emergence of new domains, the realignment of old ones, and the development of very large databases. Also, more and more, as new technologies become available, there is increasing emphasis on the retrieval of facts as opposed to the retrieval of whole documents. In particular, the Internet encourages information seeking at a micro-level while the major universal systems were designed to organize information at a macro level. Moreover, there is a growing body of research and practical application aimed at improving the situation. This papers examines three of the best known universal classification systems the Dewey Decimal (DDC), Universal Decimal (UDC) and Library of Congress (LCC) classification systems in the light of their ability to respond to the changing nature of information itself. Recent research is briefly examined for its applicability to them. Environmental science, a recently emerging domain, is used as a basis for the analysis.

2. The Problem
As “the result of historical and contemporary developments in disciplines, professions and new interdisciplinary fields” (Klein, 1996, p. 134) knowledge is in a constant state of flux. Disciplines expand, split, and spawn new sub-disciplines; new disciplines emerge and interdisciplinary and multidisciplinary relationships are forged. In the face of this, controlled vocabularies, in particular hierarchical classification schemes, rarely keep pace with the changes. This is due to several factors. New ideas and new topics appear first through communication, oral and written. At the outset, what will happen is not predictable. Some new ideas fall by the wayside; others will grow rapidly into full-fledged domains. It is only when it becomes necessary to place new knowledge into the existing scheme of things that the problem is recognized. As one author put it, one of the signs of existing disciplines breaking down is an “unease in classification” (Clark, 1996, p. 150). At the most general level, the universal systems are bound by an ancient framework of
disciplines that has its roots in the teachings of Aristotle and has been perpetuated and consolidated in academia through its curricula and the organization of its library collections. However, it is true that changes have begun to take place. For example, students can specialize in "women’s studies and obtain degrees in environment science, and programmes exist in multiculturalism, bio-engineering and numerous other cross disciplinary fields of study. Indeed there are relationships between and among subjects as yet unheard of. However, information seekers continue to struggle with the “lack of fit between interdisciplinary needs and existing knowledge, taxonomies and classification schemes” (Klein, 1996, p. 134).

3. The research record

Among the signs that some of the problems are being addressed are the many papers and conferences that have emerged in the recent past. Witness the theme titles of recent ISKO conferences that contain such words as “change” “structures and relations” and “dynamism and stability.” Response has been theoretical and practical as well as simple and complex. The idea of rearranging a scheme or relating different schemes is not new. In the 1940's the Detroit Public Library bowed to the interests of its users by creating a Readers' Interest Classification to override DDC, for the purposes of organizing browsing collections in its branches. Also, early in its history, it there were proposal to use UDC as a switching language in order to identify and relate concepts across diverse classification systems.

From the middle of the 20th century, research and development in knowledge organization has followed two distinct directions - the development of alternative approaches to existing knowledge structures, and the search for ways and means to overcome the existing knowledge structures and make them more amenable to change. With respect to alternative approaches, one result was the development of citation indexes that introduced a new method of categorizing and grouping documents and identifying hidden relationships among subjects previously unknown. The existence of these indexes then made possible a whole new research area known as literature-based knowledge discovery (Swanson, 1987; Davies, 1989). To these could be added areas of research in expert systems and artificial intelligence. In the efforts to overcome the weaknesses of the existing systems, mapping techniques were proposed to integrate the component parts of emerging domains (Olson, 1997). Multilingual tools became increasingly important (Hudon, 1997) and ultimately there has been the challenge of the Internet with its colossal size and its popular categorization of topics. How can different knowledge structures be made to communicate with each other? For example, Jacob and Albrechtsen’s (1997) research investigated the role of dialogue in unitary languages in the medical field. One of the most ambitions projects is the Integrated Academic Information Management Systems (IAIMS) with its metathesaurus, Unified Medical Languages System (UMLS). This system serves as a glossary for mapping terms to a wide range of thesauri in the medical field (Bodenreider and Bean, 2001, p. 93-94). As such, it acts as a switching language and guides the user in the use of the various databases.
4. Classification systems, integration and interdisciplinarity

With respect to changes in knowledge structures, the universal classification schemes present a particularly difficult challenge. In part, this can be attributed to the nature of the classification scheme and in part to user resistance to change because of the time, effort and costs involved. It has often been said that the ideal classification scheme is the one that is always up to date but never changes - a seemingly impossible mandate. However, as the 21st century emerges, the survival of universal classification schemes as effective retrieval tools demands that they change, adapt and accommodate access to information in the best way possible. For those who develop and revise these systems, changes raise many questions. Where should an emerging domain be placed in the overall scheme of things? What is its nature? What are its boundaries? How can it be incorporated into what already exists? If its component parts are scattered across the system how can the parts be related in a whole? If the domain is interdisciplinary, crossdisciplinary or transdisciplinary, how can the component parts be related? Can they be integrated into a whole or must they be linked across boundaries?

Clare Beghtol (1998) identifies three options for dealing with interdisciplinarity - to use existing systems with little change (i.e. remaining with the status quo); to adapt the existing general system; and to create new structures at the level of the first level of subdivision. To continue the existing system is an easy way out, but intellectually it is of dubious distinction in today’s information community. In the long run it could spell the demise of the scheme. Adaptation is attainable but the degree of success may depend on the degree of flexibility in the scheme itself. The creation of new structures has the greatest potential at the domain level because it allows freedom to create a completely new structure. Such structures may be special classification schemes in themselves. However, this is not without its problems. Newly developed structures are unlikely to be completely self contained and they may be out of sync with other parts of the system that themselves may be in need of revision and have the potential to create problems when it becomes necessary to forge links that cross boundaries. For example, in the revision of UDC 61 (Medicine) when a virus is linked to the disease caused by it, that link is between the revised class 615 and the virus in the existing class 57 class which is in need of revision. Fitting a new structure into the existing system is unlikely to be a perfect solution. (Extensions & Corrections, 2001).

Adaptation of the existing universal schemes is inevitable but there are some interesting approaches being taken in the development of new structures. The complete redevelopment of the Bliss Bibliographic Classification System (BBC2) in its second edition; the reconfiguration and expansion of UDC 61 (McIlwaine and Williamson, 1995) and the mapping of the domain of women’s studies in DDC (Olson, 1997). Beghtol’s third option, the creation of new structures at the first level of division is exemplified by the new Class 2 Religion (Extensions & Corrections, 2001) a new structure overcomes the problem of religious bias; a problem that has long plagued the Religion classes in both DDC and UDC. The revision of class 61, Medicine, in UDC is somewhat of a hybrid. It is a completely new structure but it is also an adaptation derived from another system - BBC2 class H.
5. Methodology

In the light of recent research it is important to ask how the problems of interdisciplinarity and emerging domains are being handled in three well known universal classification systems. The starting point for this investigation was the knowledge that the environmental science has clearly become an emerging domain. Moreover, its importance has also been described in an article on “The impact of interdisciplinary research in the environment sciences: a forestry case study.” (Steele and Stier, 2000). Further examination suggested that this domain was particularly suitable for this study. It is a new domain, having emerged in the latter part of the 20th century; it touches on numerous other domains (e.g. medicine, agriculture, science, etc.); and it is represented by a fast growing volume of literature. Also it has been treated somewhat differently in the three systems (DDC, UDC and LCC).

The first step in this research was to establish a frame of reference and a basis for comparison of the three systems. That frame of reference was the 4th edition of EnVoc: Multilingual thesaurus of environmental terms (formerly known as the INFOTERRA Thesaurus of environmental terms). The top terms from each of the eighteen subject categories in EnVoc were used to search each of the three systems electronically. Search systems used were “Dewey for Windows”, “Classification Plus” and “UDC Online.” The resulting data were then analyzed in more detail using the following characteristics: primary location, interdisciplinarity, degree of scatter of its component parts, and the terminology and structure of the domain. At this stage only the top terms in each of the categories were considered. As the research progresses it may be desirable to search more extensively. Ultimately, each of the three systems will be used to classify a representative set of documents chosen randomly from the an environmental science database. Finally the results of the classification will be evaluated.

6. Results

Clearly this is a work in progress and results given here represent the findings of the first stage of the research. First of all there are some differences in the capabilities of the databases searched. It is immediately apparent that the manipulation of classification schedules electronically can be very powerful when it is possible to bring up every occurrence of the term being searched whether it is a schedule caption or an example from a note. This was particularly evident in the search of DDC and UDC also provided excellent results. The least satisfactory performance came from LCC. This is, in part, a result of the kind of displays possible with Classification Plus. Several schedules can be searched simultaneously but the lack a fully integrated overall index makes the accuracy of the search problematic. Access to the LCC schedules using LCSH to identify the schedules that needs to be searched is an alternative but it is not a perfect substitute for a fully integrated index.

Secondly, an important feature is the difference in the way the domain is handled generally. In DDC21, there is no clear recognition that environmental science is a domain.

Environment in general is located in 333.7 as part of “Natural resources and energy” and this class also designated as the location for interdisciplinary works. Here also provision is made for further subdivision under “general topics” such as
“environmental impact studies” and ‘management and control’. The UDC International Medium Edition, English Text, on the other hand has clearly identified a prime location for “Environmental science, Environmentology” in 504. This places it logically at the beginning of Class 5 and provides it with its own set of subdivisions; the primary subdivisions being the type of environment (e.g. 504.3 Atmospheric environment; 504.4 Hydrospheric environment, etc). Both systems scatter aspects of the topic but UDC provides the ability to bring these aspects together at 504 through its liberal provisions for combination whereas DDC is less flexible to this possibility. With LCC general works on the environment appear to be located in QC with some subtopics scattered through other classes but little chance of crossing boundaries. All three systems show scatter of the component subtopics of environment but UDC is the only system that provides the ability to bring all aspects of the domain together physically. The use of the colon to link concepts from across the system is a major contribution to its flexibility. Further analysis of the primary sites will take place later in the research.

When the terms from the 18 categories of the domain were searched, the most striking characteristic revealed was the extent of this domain and how widely its components are spread across the schemes. For example, in UDC, the term “housing” appeared in 53 records ranging across class numbers in the 3, 6 and 7 categories, The term “management” occurred in UDC 157 times ranging from 004 to 796. In the search of all three tools all of the terms were located with minor exceptions. In some cases, terms took different forms. For example, the term “lithosphere” was located in DDC; in UDC it appears as “lithospheric environment.” Also, the term “human settlements”, a term familiar in United Nations information resources, was used as a category in EnVoc. This term was not found in either UDC or DDC but it was located in LCC. “Housing” seemed to be the alternative, and DDC and UDC yielded a goodly number of records in both cases (DDC, 68 hits; UDC, 53 hits). Thus it appears that there may be some surprises in terminology and some important differences. Not surprisingly, it is abundantly clear from these preliminary searches that the scheme that makes the greatest use of facet analysis is likely to be the most accommodating (i.e. UDC).

UDC’s use of the colon to join class numbers from across the scheme is important to flexibility. While UDC, is out of date in some areas, this is gradually being rectified. “Environmental Science” (504) was established as a domain in the second edition of UDC’s International Medium Edition, English Text and provides an excellent location for gathering the parts of the domain. Class 333.7, DDC’s counterpart, is workable but less logically developed.

6. Further study

Clearly this is a work in progress. This preliminary study is only the beginning of this research. More in-depth analysis is needed as a next step. Continuing with the data collected, part of the investigation will be to examine the terminology and the structure involved and to assess the ability of each system to handle retrieval successfully.

References
Beghtol, Clare. (1998). General classification schemes: Structural principles for multidisciplinary specification. In Structures and relations in knowledge organization:


M.V. Hurtado, L. García and J. Parets
HTS Computer Engineering, University of Granada, Granada, Spain

Semantic Views over Heterogeneous and Distributed Data Repositories: Integration of Information System Based on Ontologies. 1

Abstract: Nowadays, in the global information infrastructure, the great expansion of the communication networks allows users to work with a great quantity of heterogeneous and distributed data, which are stored in different repositories. In order to support co-ordinated access to this data, preserving the correct semantics of the data stored in different ways, an integrated view is needed. We propose, in a first step, obtaining an integrated schema (Semantic View) and an integrated terminology in order to describe the content of the repositories. Ontologies are suggested as an interesting semantic representation which can be used here. Next, in a second step, the integrated schema and integrated terminology must be enriched in order to allow us to manage the intrinsic evolutionary aspects which characterise the modelling process of global information systems.

I. Introduction
In recent years the development of the communication networks allow users to work with a great quantity of data repositories. These repositories present different structures, organisations, query languages and data semantics. A partial solution to the problem of lack of uniformity when dealing with the available repositories consists on defining new information retrieval techniques with a strategy that focuses on information content and semantics.

In the second section, we will describe some reasons that justify the use of Semantic Views over the repositories. In the third section, as a result of research carried out in the last decade, Ontologies are suggested as an interesting semantic representation. Following the approach of previous papers on software evolution (Parets and Torres, 1996), (Parets et al., 1999) and (Rodriguez et al., 2000) and on information integration (Hurtado and Parets 2001), the fourth section of the paper presents the possible evolution mechanisms of the semantic views. Additionally, effective graph mechanisms for semantic evolution are provided.

2. Semantic Views: Implications and Needs
Semantic Views on specific domains minimise the problem of semantics associated with stored data and hide all the technical and organisational details associated with data. Furthermore, a logical schema of shared information is obtained which provides a logical description of the shared data, allowing application programs and databases to interoperate without having to share data structures. Indeed, in order to provide definitions for the vocabulary used to represent knowledge, a richer semantically integrated terminology is obtained too. In order to do this, the formalisms used for Semantic Views must allow:
A common understanding of the structure of information among people or software agents.
- The reuse of domain knowledge.
- To make domain assumptions explicit.
- A separation of domain knowledge from the operational knowledge.
- An analysis of domain knowledge.

Typically, Semantic Views can be described by means of different formalisms (systems based on Description Logic (Devanbu, 1994)), terminological systems, ontologies and so on).

We advocate here for using ontologies, because they have moved beyond the domain of library science, philosophy and knowledge representation. Ontologies are used as central controlled vocabularies that are integrated into catalogues, databases, web publications, knowledge management applications, etc. and large ontologies are essential components in many online applications including retrieval (such as Yahoo and Lycos) and e-commerce (Such as Amazon and eBay) (McGuinness et al., 2000).

3. Using Ontologies

In the context of knowledge sharing, the term ontology is used as an explicit specification of a conceptualisation (Gruber 1995). People typically have some notion of the meaning of the term, one of the simplest notions of an ontology may be a controlled vocabulary (i.e., a finite list of terms). Catalogues are an example of this category, they can provide an unambiguous interpretation of terms. Another potential ontology specification is a glossary (a list of terms and meaning) and also a thesaurus, because they provide additional semantic relationships among terms (such as synonym or hierarchical relationships).

Furthermore, because ontologies have to express more information about shared conceptualisation and also conceptual frameworks for modelling domain knowledge, we will require that an ontology holds the following properties in order to consider something as an ontology:
- Finite, controlled and extensible vocabulary
- Unambiguous interpretation of class and term relationships
- Hierarchical subclass relationships between classes
- Specification of arbitrary logical relationships between terms.

In this sense, we propose a specialised kind of ontologies based on the Object-Oriented Approach which satisfies the previous properties. In the O-O approach, classification is a core concept (Rumbaugh, 1991). The objects which share data structure and behaviour are grouped into classes, and the classes can be sub-classified by means of inheritance mechanisms (i.e. in figure 1, Author denotes a class and the class definition in natural language can be the following "An author is a person who writes things. An author must have created at least one document. In this ontology, an author is known by his or her name"). This implies a high degree of abstraction, which describes important properties and ignores irrelevant ones, and this is a conceptual process which is independent of the programming languages (i.e. the sentence gives necessary conditions on class membership, says that authors must also be persons or says that authors must have exactly one associated name. This means that the relation author.name maps every instance of
the class Author to the same name, with only one name per author). This process is usual in human knowledge acquisition and allows abstract concepts to be expressed.

The classification schema in O-O is always explicit, for instance, figure 1 shows a partial enriched view of the possible UML class diagram based on the Stanford-I ontology (http://www.ksl.stanford.edu) about bibliography data which can be considered as special ontology. In addition O-O classification implies a description of the static structure of the classes of the system and their relationships (associations, specialisation, generalisation, etc).

![Figure 1. The proposed, UML class diagram based on Stanford-I ontology about bibliography data](image)

Previous characteristics increase maintainability in ontologies based on the object model because:

- Explicit classification facilitates the introduction of new classes and the re-structuring of previous ones.
- Inheritance allows code reusability (the code of the behaviour of a class is reused by its subclasses), code sharing (different users and systems can share the same classes), interface consistency (inheritance guarantees that the inherited behaviour is the same for the subclasses and that the objects of the subclasses interact in a very similar way) and rapid prototyping (the classes developed in previous systems can be reused and refined).
- Information hiding allows the code of the behaviour of a class to be changed without changes in the uses of the class.

4. **Evolutionary aspects in the integrated schema**

The use of semantic views based on Ontologies allows, on the one hand, the flexible encapsulation of data in repositories and, on the other hand, the sharing of information between different users or systems. Many disciplines develop standardised ontologies that domain experts can use to share and annotate
information in their fields. An ontology defines a common vocabulary for researchers who need to share information in a domain. It includes machine-interpretable definitions of basic concepts in the domain and relations among them. We have proposed that ontologies be described using an object-oriented approach and therefore using graphical representation and graph restriction.

This paper shows the application of our experience in graph evolution to semantic views, specially the idea of defining semantic restrictions over the relationships represented in the graph. Finally, once again our approach provides mechanisms to propagate changes, and it focuses on the evolutionary nature of Global Information Systems.

There is an additional and important fact that cannot be forget, the intrinsic evolutionary aspects, which characterise the modelling process of global information systems. That is to say changes in the underlying repositories or in the Semantic View may produce a propagation of changes in the rest of the global information system.

A previous paper (Hurtado&Parets2001) shows that certain advantages in the specification of the evolution of semantic views can be obtained when advanced object-oriented techniques, such as dynamic classification, multiple inheritance, etc. is used. Figure 2 shows that the class Proceedings can maintain inheritance relationships with Document and Conference, this fact it's not allowed in classical formalisation of ontologies (Guarino, 1998)).

These techniques include the notions of identity, classification, polymorphism and inheritance, which provide interesting ways of organising the objects and their activity and improve the evolutionary nature of semantic views.

Moreover, in order to provide concrete mechanisms of change, we propose that the previous ontology (figure 1) can be represented as heterogeneous graph where the nodes are classes (ellipses in the graph) and arcs are different types of relationships between them (associations, specialization, generalization, etc.) or relationships defined by users (write, Read-in etc) (see figure 2).

Additionally, this kind of graph has not only special semantics adopted from O-O relationships but also some implicit and explicit restrictions for these relationships (for instance, the relationship KindOf always verifies some restrictions such as a-cyclic restriction, anti-symmetric property, and it is incompatible with PartOf and with Is-a). The relationships defined by the user and their restrictions can also be explicitly defined.

We are now researching on practical ways of specifying these graphs. In previous papers (García and Parets,2000) and (García-Cabrera et al.,2001) we presented a set of basic operations on the graph (such as Create_node, Del_node, Create-arc, Del_arc, Nodes-connected-to, Connection-by, etc.) and a set of basics restrictions on the graph (tree, acyclic_graph, weak_connected, etc.), restrictions on the nodes (i.e. each node must have a label, unique name assumption, etc.) and restrictions on the arcs (have a label, anti-reflexive, anti-symmetrical, transitive, a-cyclic, incompatible with, incompatible transmission, and so on).

These operations allow that the structure of the graph to be changed, and the set of restrictions on the graph, nodes and arcs helps in propagating the changes because, the preconditions of each operation must be checked before a change is made. This fact implies that propagation of changes can be represented and automated.
For instance, figure 3 shows how the semantic view evolves as a result of introducing new classes and specialisation relations among them, this fact causes, for example, a change propagation in the mapping rules. Figure 3 shows the new class Best-seller which is a Trade-Book with more than 1000 copies sold and the prohibition of adding a new KindOf relation between Book and Miscellaneous-Publication as result of the acyclical restriction. More complex changes, as the introduction of intermediate new classes or the deletion of classes could be introduced.

In order to make changes in the integrated information, a graphical approach has important advantages. From a formal point of view, to check the preconditions of operations and restrictions on the graph is easier for propagation changes. In addition abstract data types or classes can be used for implementation. Therefore, from a user point of view, this approach is easier to understand because it encodes the shared information by means of a simple graph with the labelled concepts, relationships and restrictions.

5. Conclusion

The use of semantic views based on Ontologies allows, on the one hand, the flexible encapsulation of data in repositories and, on the other hand, the sharing of information between different users or systems. Many disciplines develop standardised ontologies that domain experts can use to share and annotate information in their fields. An ontology defines a common vocabulary for researchers who need to share information in a domain. It includes machine-interpretable definitions of basic concepts in the domain and relations among them. We have proposed that ontologies be described using an object-oriented approach and therefore using graphical representation and graph restriction.

This paper shows the application of our experience in graph evolution to semantic views, specially the idea of defining semantic restrictions over the
relationships represented in the graph. Finally, once again our approach provides mechanisms to propagate changes, and it focuses on the evolutionary nature of Global Information Systems.


**Figure 3 Changes in the graph**

**Notes**
1. This research is partially supported by a R+D project of the Spanish CICYT (TIC2000-1673-C06-04).

**References**


Abstract: The recent transformation of separate economic classification schemes in Canada, Mexico, and the United States into a single North American system makes it possible to collect and organize data across these nations' boundaries. In reconstructing the development of the North American Industry Classification System (NAICS) as it coincided with the creation of the North American Free Trade Agreement (NAFTA), we foreground the fundamental roles of economic theory and political context. Although it remains to be seen how effective NAICS will be as a component of an information infrastructure designed to support a new regime of production and consumption, it is clear that the system makes possible the transnational data collection and analysis that will shape future understanding of the NAFTA-generated space of economic activity.

1. Background of NAICS

In the past decade, the United States government transformed its industrial classification scheme for the organization of economic data from a purely national to a supranational endeavor involving Canada and Mexico. The change coincided with the three countries' signing of the North American Free Trade Agreement (NAFTA) and reflected their interest in facilitating the integration of business-related knowledge across national boundaries in the new NAFTA-generated spaces of production. The result is the North American Industry Classification System (NAICS), which divides the economy into twenty broad "Sectors" designated by two-digit class numbers and then into finer distinctions down to the five-digit level for transnational comparability and down to the six-digit level for country-specific data. NAICS replaces the U.S. Standard Industrial Classification (SIC), Canada's Standard Industrial Classification, and Mexico's Clasificación Mexicana de Actividades y Productos.

The introduction of NAICS signified a reconceptualization of the economy as a global rather than local phenomenon. While it is still unclear exactly who in the government or the private sector will take advantage of the potential for international comparability that NAICS represents, it is obvious that the system is a key component of the "information infrastructure" (Star & Ruhleder, 1996) that supports a globalizing move toward free trade within the North American region. Such infrastructural components do their work behind the scenes and are seldom noticed by those who use the information systems they support (Bowker and Star, 1999, p. 2-6). Consequently, in this paper we use selected government documents to understand how overlooked but crucial information organization schemes come to be created and implemented. We also explore the fundamental role that economic
theory and political context played in the emergence of NAICS, at the expense of classificatory theory. Finally, we touch on the fact that schemes for organizing data about industries serve an essentially different purpose than schemes for organizing library materials.

2. Economic theory and the emergence of a single principle of classification

Although officials from Canada, the U.S., and Mexico were negotiating specific features of NAFTA throughout 1991 and 1992, those negotiations did not have an immediate impact on discussions within the U.S. regarding the need for a new economic classification system. In November of 1991, the U.S. Bureau of the Census hosted an “International Conference on the Classification of Economic Activities” in the Washington, D.C. area. About 40 individuals from U.S. and other government agencies, universities, and the private sector participated. Two officials from Statistics Canada, that nation’s centralized agency for data gathering and reporting, were on the conference program, but no one from Mexico’s comparable agency, Instituto Nacional de Estadística, Geografía e Informática (INEGI), appears in the proceedings (U.S. Bureau of the Census, 1992).

The concerns of economists and statisticians from U.S. government agencies, especially the Bureaus of Economic Analysis, the Census, and Labor Statistics, dominated the conference. Conference organizers from the Census Bureau were intent on understanding what should be done about the outdated SIC within the U.S. government that employed them; no explicit call for the U.S. to cooperate with its neighbor nations is apparent in the conference proceedings. The name reflected the reality: It was an international conference, not a conference on international classification nor on classification of global economic activities.

A few key themes did emerge during the conference. One was the belief that the U.S. economy had changed so significantly during the post-industrial latter half of the twentieth century that the SIC system, originally created in the industrial 1930s, even with its several revisions and updates over time, was hopelessly antiquated. The other factor was the belief that the old system employed an incoherent and shifting set of criteria in the creation of new and revision of existing coding categories.

At the 1991 conference, Bureau of Economic Analysis Chief Economist Jack Triplett (1992a) asserted the need to draw on economic theory, not other coding methods, to inform the structure of the new system. Along with a few other conference participants, Triplett also mentioned the idea of a system that broke away from SIC’s hierarchical structure. Speakers who addressed the issue of hierarchy recognized that new information technologies made it possible to have a dynamic, non-hierarchical system that would allow users to combine the data into groupings that served their individual needs (Gort, 1992; McGuckin, 1992). Revealing his focused perspective, Triplett (1992a, 1992b) suggested that a hierarchical structure might not be needed; it depended on the underlying organizing principle derived from economic theory that was chosen as the basis of the new system. Not until that was decided could the question regarding a hierarchical structure be decided. Others who addressed the issue came to a somewhat different conclusion. They assumed the new classification system would be hierarchical in its representation of the economy (Popkin, 1992; Young, 1992; Issues Paper No. 2, 1993). The difference for these speakers lay in whether the
hierarchy were created from the top down, in an idealized picture of the economic structure, or from the bottom up, in a pragmatic approach to creating only what was needed.

For researchers from the library and information science discipline, it is tempting to perceive the question of a top-down hierarchy as analogous to Dewey Decimal Classification and a bottom-up approach as in sympathy with Library of Congress Classification. While library classification is concerned with maintaining and supporting the distinct individuality of a work, however, economic classification, as Desrosières (1998) notes, is concerned with the aggregation of individual items and the consequent erasing of their individual uniqueness.

Classification systems serve a collocation function by grouping together items that are similar in some way. In NAICS the point of similarity encompasses "like production processes" or "inputs," rather than products or outputs. That is, business establishments that utilize the same kinds of workers, raw materials, and tools are grouped together. This notion of "like production processes" is the economic principle underlying NAICS as a hierarchical structure and a coding system (Issues Paper No. 1, 1993). One can look at the NAICS organizational structure and see, for example, that the system's creators believe that "performing arts companies" and "independent artists, writers, and performers" deploy sufficiently different production processes to be separated into distinct categories (7111 and 7115, respectively) but sufficiently similar to be grouped together in the larger 711 subsector and ultimately in Sector 71. (See Fig. 1) This grouping, or collocation, phenomenon functions as well in library classification. In economic classification, however, collocation is a consequence of information organization rather than its main purpose. In economic classification, aggregation of data is the main purpose.

This is important in the context of a globalizing classification system such as NAICS because of the issue of comparability. Canada, Mexico, and the U.S. have agreed that each level of NAICS categories will be comparable from the topmost two-digit level to the five-digit categories. They also have agreed that each nation can create its own six-digit categories for finer gradations appropriate to the particular characteristics of each economy, but that these categories must not affect the aggregation of data at the other four levels that support cross-border comparability (1997 North American industry classification system, 1997). It is not a question of whether each nation is entitled to a more elaborate classification hierarchy than is available in their shared system. It is a question of whether the data add up in a way that preserves the underlying intent behind comparability within the North American region as one vast economic zone. In time, this zone will produce the data that, when collected within the categories of NAICS, will confirm the concreteness of the new economic space.

3. From national to supranational classification

At the 1991 conference and in subsequent work by the ECPC, the topic of aggregation was central to the ideas surrounding a new classification scheme. International comparability was not. Of six issues papers disseminated by the ECPC, one covered the topic of aggregation (Issues Paper No. 2, 1993). The paper discussed aggregation not in the context of international comparability but in the context of hierarchical structures of classification schemes. Much of it recounted
points made at the international conference in 1991. For example, the issues paper noted that William Seltzer of the United Nations Statistical Office spoke at the conference "[F]rom an international vantage point," but quoted him not on the topic of international comparability but on the collocation and aggregation functions of classification. Organization of statistical data was "first, a system for grouping units in an ordered way for the purpose of data collection and storage, and second, a hierarchical system of grouping units for the purpose of aggregation, tabulation and analysis" (qtd. in Issues Paper No. 2, 1993, 17001).

Despite the ECPC's lack of interest in international comparability, as evidenced in the content of its issues papers series, the idea came up in a separate document, the ECPC's report summarizing the public comment received on its first two issues papers (Summary of public comments, 1993). In that report, the ECPC acknowledged that the idea for creating a classification system that would work across the NAFTA region about to become official on January 1, 1994, came from outside the U.S. government and its statistical agencies. (The ECPC also announced in this paper that it would issue a seventh issues paper on international comparability, but it never did; the series stopped with issues paper No. 6.) The paper quoted only two interested parties making such requests. One was a representative of a trade organization, the Manufacturing Jewelers and Silversmiths of America, who noted the development of a global economy. The other, representing the Steel Service Center Institute, requested comparability across North America because its members came from the entire region.

The ECPC followed up quickly. By July of 1994, the Office of Management and Budget (Standard Industrial Classification Replacement, 1994) was able to announce that the three NAFTA signatories had a work plan in place that would make it possible to implement the system when the next U.S. economic census was to occur, in 1997. A year later, OMB began issuing the first of 31 agreements negotiated among Canada, Mexico, and the U.S. regarding the detailed structure and nomenclature of NAICS. Having initiated a sweeping reconsideration of its internal economic classification system, U.S. statistical agency officials were slow to realize that the debates and discussions surrounding the creation of NAFTA had an impact on their work as well. Once industry stakeholders called their attention to the significance of transborder data organization, they involved their counterparts in Mexico and Canada in the creation of a classification scheme acceptable to all concerned.

The emergence of the transnational character of NAICS brings to mind Alain Desrosières' (1998) observation that the work of encoding and measuring is shaped within a particular social and political context which it responds to and which it in turn helps to shape. In the case of NAICS the work of encoding became closely associated with the new regime of production and consumption that was made concrete with the signing of the NAFTA agreement. NAICS was repositioned not just as a radical new mode of industrial classification but also as the structure within which information can be collected about an economic region that transcends the borders of the United States.

4. Conclusion

We have argued and illustrated that NAICS is a major classificatory project, yet the development of two of its most distinguishing characteristics--a single
organizing principle underlying the scheme and its supranational character—relied little if any on traditional theories of classification. In the case of the organizing principle the authors of NAICS resorted to economic theory and in the case of internationalization the authors responded, almost as an afterthought, to the emerging context of NAFTA. NAICS is still evolving, with new changes being implemented in 2002. It remains to be seen how effective NAICS will be as a component of an information infrastructure designed to support a supranational understanding of economic activity. Each of the three nations has issued its own NAICS manual indicating where the categories match exactly with one or both of the other two countries and where they diverge (Mexico Instituto Nacional de Estadística, Geografía e Informática, 1997; Statistics Canada, 1997; U.S. Office of Management and Budget, 1998). It is difficult to predict how economic data coded according to NAICS will be used, whether by governmental agencies, non-governmental organizations, or private-sector analysts. Will it be for comparability of production processes across the three nations or for gauging total productivity by sector within the NAFTA region? Will it preserve or erase national boundaries in the context of free trade and globalization? What is certain with NAICS is that the enabling mechanism for supranational data classification and analysis is in place.

71 Arts, Entertainment, and Recreation
711 Performing Arts, Spectator Sports, and Related Industries
7111 Performing Arts Companies
71111 Theater Companies and Dinner Theaters
71112 Dance Companies
71113 Musical Groups and Artists
71119 Other Performing Arts Companies
7112 Spectator Sports
71121 Spectator Sports
711211 Sports Teams and Clubs
711212 Racetracks
711219 Other Spectator Sports
7113 Promoters of Performing Arts, Sports, and Similar Events
71131 Promoters of Performing Arts, Sports, and Similar Events with Facilities
71132 Promoters of Performing Arts, Sports, and Similar Events without Facilities
7114 Agents and Managers for Artists, Athletes, Entertainers, and Other Public Figures
71141 Agents and Managers for Artists, Athletes, Entertainers, and Other Public Figures
7115 Independent Artists, Writers, and Performers

References


An Exploratory Study of Human Clustering of Web Pages

Abstract: This study seeks to find out how human beings cluster Web pages naturally. Twenty Web pages retrieved by the Northern Light search engine for each of 10 queries were sorted by 3 subjects into categories that were natural or meaningful to them. It was found that different subjects clustered the same set of Web pages quite differently and created different categories. The average inter-subject similarity of the clusters created was a low 0.27. Subjects created an average of 5.4 clusters for each sorting. The categories constructed can be divided into 10 types. About 1/3 of the categories created were topical. Another 20% of the categories relate to the degree of relevance or usefulness. The rest of the categories were subject-independent categories such as format, purpose, authoritativeness and direction to other sources. The authors plan to develop automatic methods for categorizing Web pages using the common categories created by the subjects. It is hoped that the techniques developed can be used by Web search engines to automatically organize Web pages retrieved into categories that are natural to users.

1. Introduction

The World Wide Web is an increasingly important source of information for people globally because of its ease of access, the ease of publishing, its ability to transcend geographic and national boundaries, its flexibility and heterogeneity and its dynamic nature. However, Web users also find it increasingly difficult to locate relevant and useful information in this vast information storehouse. Web search engines, despite their scope and power, appear to be quite ineffective. They retrieve too many pages, and though they attempt to rank retrieved pages in order of probable relevance, often the relevant documents do not appear in the top-ranked 10 or 20 documents displayed. Several studies have found that users do not know how to use the advanced features of Web search engines, and do not know how to formulate and re-formulate queries. Users also typically exert minimal effort in performing, evaluating and refining their searches, and are unwilling to scan more than 10 or 20 items retrieved (Jansen, Spink, Bateman & Saracevic, 1998).

This suggests that the conventional ranked-list display of search results does not satisfy user requirements, and that better ways of presenting and summarizing search results have to be developed. One promising approach is to group retrieved pages into clusters or categories to allow users to navigate immediately to the “promising” clusters where the most useful Web pages are likely to be located. This approach has been adopted by a number of search engines (notably Northern Light) and search agents.

But what kinds of categories are likely to be useful to the user and help the user locate relevant Web pages? Perhaps the useful categories are those that the users would spontaneously use in grouping or clustering the Web pages. This
exploratory study seeks to find out how human beings cluster Web pages naturally. Twenty Web pages retrieved by the Northern Light search engine for particular queries were presented to subjects, who were asked to sort the pages into groups that were natural or meaningful to them and that were likely to be useful to help them locate the pages they wanted. They were also asked to give a descriptive label to each group.

This study sought to answer the following questions:

1. What kinds of clusters and categories are created, and how similar are the clusters created by different people?
2. What criteria do people use to decide on the categories and to assign Web pages to categories?
3. Are there "universal" or common categories that are created by many users?
4. Are there differences between the clusters and categories constructed by subjects who contributed the query and subjects who did not contribute the query?

Our expectation was that many of the categories formed will not be subject-related or topical categories, but pertain to the form of the documents, the purpose of the author or the type of treatment given to the subject—and other aspects that cut across subject categories and are, in that sense, universal.

In this paper, the terms cluster and category are used interchangeably. A cluster is a subset of Web pages that the subject considers to be similar in some way and belong together. A category is a descriptive label assigned to a cluster by the subject, as well as the concept used to represent the cluster. The assumption is that human clustering involves the construction of a mental concept or category to which Web pages are assigned.

While there has been a substantial amount research on automated methods for clustering related Web pages, there is a paucity of research on human clustering of Web pages. Macskassy, Banerjee, Davison & Hirsh (1998) performed an exploratory study of human clustering of Web pages with 10 subjects who each sorted the Web pages retrieved for 5 queries. They found no discernible pattern in how different subjects clustered Web pages. The inter-subject similarity of the clusters created was low—an average similarity of 0.28 for subjects who were given only the URLs and titles of the Web pages, and an average similarity of 0.16 for subjects who were given the full-text of the Web pages. Subjects generally created small clusters, and those with access only to URLs and titles created fewer clusters than those with access to the full Web page. When given the full-text of Web pages, the overlap between clusters increased, suggesting that clustering methods that permit overlap are more appropriate than those which do not. Subjects were found to display different behavior across queries in terms of cluster overlap, and number and size of clusters. The study has a major limitation in that the Web pages used were retrieved by the Rutgers University WebWatcher search engine with a much smaller database and narrower subject scope than regular search engines.

2. Research Method

In this study, ten colleagues and friends of the authors were asked to contribute a search query that they had recently submitted to a Web search engine. The queries collected are quite diverse, and are as follows:
3. Results

3.1. Number and size of clusters
Table 1 shows the number of clusters created by each contributor and non-contributor for the queries. Overall, the subjects created an average of 5.4 clusters for each sorting. The average size of a cluster was 3.9 Web pages. There were few overlaps between clusters, i.e. subjects rarely assigned a page to more than 1 cluster.

The number of clusters created varied from 3 to 9, but the average number of clusters was stable across the queries, ranging between 4.3 and 5.7—except for query Q9 (Roman occupation of Britain) with an unusually high average of 7.7.

This suggests that people have a common sense about how many clusters are appropriate, and tend to create the same number of clusters for different queries. It was observed that subjects were conscious of the need to avoid creating too many
or too few categories and categories that were too big. When subjects felt that they were assigning an increasing number of pages to one category, they began to consider whether the category should be split into more categories.

However, some people tend to form more clusters than others. Contributor no. 7 (who created 6 and 7 clusters) and contributor no. 9 (7 and 9 clusters) each created the highest number of clusters for the two queries that each sorted. On the other hand, contributor nos. 1, 4 and 5 created 3 or 4 clusters per query, and created the lowest number of clusters for the queries they sorted.

Query Q9, which obtained a high number of clusters, yielded more topical categories than other queries. Because of the nature of the topic and the Web pages retrieved, the subjects (especially the contributor of the query) could easily discern sub-topics covered by the different Web pages.

Cluster sizes vary substantially, with many small-sized and big-sized categories. The 30 sortings performed in this study (3 sortings per query) yielded a total of 37 singleton clusters and 19 clusters of size 8 or larger.

3.2. Similarity of clusters

We calculated the inter-subject similarity among the three sets of clusters created for each query, using the similarity measure employed by Macskassy et al. (1998). It is based on the frequency with which the two subjects assigned common pairs of pages to the same cluster. To determine the similarity between two sets of clusters (i.e. two sortings), set 1 and set 2, all possible pairs of Web pages in each cluster are identified, i.e. we obtained the set of same-cluster-pairs of pages for set 1 and for set 2. We then determine how many pairs are common between set 1 same-cluster-pairs and set 2 same-cluster-pairs, and the total number of unique pairs in the union of set 1 same-cluster-pairs and set 2 same-cluster-pairs. The similarity between the two sets of clusters is calculated using the formula:

\[
\text{Sim} = \frac{|\text{set 1 same-cluster-pairs} \cap \text{set 2 same-cluster-pairs}|}{|\text{set 1 same-cluster-pairs} \cup \text{set 2 same-cluster-pairs}|} = \frac{\text{no. of common pairs in set 1 and set 2 same-cluster-pairs}}{\text{total no. of unique pairs in the union of set 1 and set 2 same-cluster-pairs}}
\]

The overall average inter-subject similarity was a low 0.27. This is higher than the similarity of 0.16 obtained in the study by Macskassy et al. (1998). The average similarity for each query ranged from 0.15 to 0.37. We calculated the average similarities between

- contributor of the query versus contributor of another query
- contributor versus non-contributor
- contributor of another query versus non-contributor.

The three average similarities were about the same.
Table 1. Number of clusters created by contributors and non-contributors for each query

Note: The values in italics refer to the clusters created by the contributor of the query. These values occupy the diagonal of the table.

<table>
<thead>
<tr>
<th>Query</th>
<th>Average per contributor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>Q2</td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>Non-contributors</td>
<td></td>
</tr>
<tr>
<td>Average per query</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3. Nature of categories

We found, as did Macskassy et al. (1998), that different subjects clustered the same set of Web pages quite differently and created different sets of categories. Across different sortings of the same Web pages, only 1 to 3 categories appeared more than once. However, the categories created can be divided into the following types:

1. **Topical categories**, usually sub-topics of the query topic (52 categories)
2. **Degree of relevance**, e.g. "relevant", "some relevance", "irrelevant", "can’t decide if it’s relevant" (17 categories)
3. **Degree of usefulness**, e.g. "useful information", "most useful pages", "potentially useful", "less useful", "potential leads", "may be useful", "useless", "totally useless", "redundant", "follow up if nothing else is useful" (16 categories)
4. **Authoritativeness of the source**, e.g. "authoritative pages", "pages set up by personalities" (2 categories)
5. **Depth of coverage or length of page**, e.g. "brief information", "basic information", "general information", "introduction", "definitions", "background information", "detailed information", "short articles", "long articles" (15 categories)
6. **Accessibility criteria, directory of sources or direction to other sources**, e.g. "links", "bibliographies", "further resources", "Web resources", "books on ...", "articles on ...", "pages I’ve to purchase before viewing", "discussion groups", "search engine" (26 categories)
7. **Format or genre**, e.g. "lecture notes/slides", "articles", "case studies", "biographies", "news items", "personal anecdotes", "technical papers", "chronology" (14 categories)
8. Purpose of the author or target audience, e.g. “educational …”, “travel guides”, “FAQs”, “advertisements”, “companies selling products/services”, “useful for students”, “educators would find these helpful”, “guide for businessmen” (11 categories)

9. Language, e.g. “non-English”, “foreign language” (5 categories)

10. Number of occurrences of an important feature, e.g. number of images, number of hotels (4 categories)

In parentheses are given the total numbers of categories created by the subjects that fall under the particular type. The distinction between format (item 7) and purpose (item 8) is fuzzy, since document format and the author’s purpose are often related. Most subjects used multiple types of categories within each sorting.

Nearly 1/3 of the 162 total number of categories created were topic-related categories, usually sub-topics of the query topic. Subjects’ think-aloud verbalizations suggest that people with more knowledge of the subject area of the query are more likely to create topical categories. For query Q1 involving cartoon wallpapers, subjects with more knowledge of the domain created the topical categories of “Japanese cartoons”, “Western cartoons”, “Oriental cartoons”, etc., whereas the subject with little domain knowledge created the categories of “single image” and “multiple image” categories. There is some indication that the contributor of a query is more likely to create topical categories because the contributor is likely to have more domain knowledge. Whether topical categories are created, and what the categories are also depend on the topic of the query. Broader topics seem to yield more topical categories because the sub-topics are more obvious to the subject. However, these conclusions are very tentative because of the small sample size.

When asked about the strategy the subject used in sorting the Web pages, many subjects made references to “subject”, “topic”, “content” and “focus”. In reality, many non-topical categories were constructed. No subject created exclusively topical categories, and every sorting yielded some non-topical categories. Some subjects categorized Web pages into different degrees of relevance and usefulness. Contributors of the queries seem more likely to construct this kind of categories. Non-contributors exhibit some reluctance in making relevance judgments. One subject categorized the Web pages for Q3 (Hepatitis B and liver cancer) by the authoritativeness of the source. She considered this criterion important because of the medical nature of the query.

We observed that during clustering, the subjects tended to first identify the most salient attribute or dimension of the Web page. The attribute need not be topic-related. This salient attribute appears to be the main criterion used in assigning the page to a category. So, Web pages are typically categorized based on a single dimension. This is in line with the findings of cognitive psychology research that people have a strong tendency to create categories based on a single dimension (Medin et al., 1987). Different dimensions are used for different clusters, but how the subject selects a dimension as salient is not known.

We hypothesize that each person has a mental library of prototypical Web pages that exemplify the different dimensions. During categorization, the subject identifies the prototype that is closest to the Web page being categorized. The dimension associated with this closest prototype is then selected as the salient dimension of the Web page. Whether such mental prototypes of Web pages exist, what they are, how the similarity between a Web page and prototype is calculated,
and whether some dimensions or prototypes are weighted differently in different circumstances can be the subject of future study.

4. Conclusion

This study has found that different people cluster Web pages differently for the same set of Web pages, and construct different sets of categories. Inter-subject similarity of the clusters created was low. The categories constructed can however be divided into 10 common types. About 1/3 of the categories created were topical, another 20% of the categories are about the degree of relevance or usefulness. The other categories relate to subject-independent categories such as format, purpose, authoritativeness and direction to other sources. These categories are applicable to any topic and query.

We plan to develop automatic ways of categorizing Web pages into these subject-independent categories. We also hope to identify general topical categories that are likely to be applicable to many queries, and then develop automatic methods of assigning Web pages to these categories. We need to test our hypothesis that organizing Web search results into these categories do help users identify useful Web pages more effectively and efficiently. Finally, we hope to carry out a more in-depth study of the cognitive processes involved in human clustering of Web pages.

References


Stéphane Chaudiron, Majid Ihadjadene and François Role  
Department of Information-Communication (lphi)  
University of Paris10, France

Authorial Index Browsing in an XML Digital Library

Abstract: In this paper, we present the CODEX project which aims at designing a system for ETDs access based on a XML architecture. XML offers a possible delivery format for controlled vocabularies. It can also be used as an alternative to offer a single interface to the different controlled vocabularies and classification systems. We use these XML facilities to implement a system based on the authorial indexes as a browsing tool.

1. Introduction

The widespread use of the Web among the scholarhip community and the increasing availability of thesis as electronic documents raise new opportunities for a wider dissemination of large scientific documents in electronic form. The true benefit of an ETD (Electronic Thesis and Dissertation) digital library, however, comes not from the replication of traditional library functions, but rather from the ability to make possible tasks that would not be possible outside the electronic environment, such as the hypertextual linking of related texts, full text searching of holdings, and the integration of filtering information tools and data visualisation. Large corpora of ETDs on the Web also gives the researchers a new environment to consider the question of knowledge classification and organization. As a matter of fact, the question of information structuration within the ETDs and, specifically, the matching process between the document produced by the author and the informational need that may have a reader has been renewed by the use of new techniques of classification and description.

Different strategies are available to access the information contained in ETDs. Classical — even not easy to implement — ones are based on classifications, thesaurus or ontologies which give the user (i.e., the reader) keys to go in the informational nodules at different levels of granularity. These strategies share the fact that they all are pre-defined views of a knowledge domain. Pre-defined views are also tools which are all conceived by library and information science professionals who share some common concerns — at least practical if not cultural — about what is a representation. In this context, readers need to access the information in the documents via these pre-defined representation tools.

Another way to consider the matching process between the document and the reader is to examine what happens when using the author-made index instead of the pre-defined representation tools. The author abstract and index (or indexes) may be considered as content representation tools with specific features. Authorial indexes directly classify the ETDs according a personal view (the author one) of the content and give the reader a direct access to the document. One may consider that when indexes are directly produced by the authors, documents are faster available on the Web and at a reduced cost. Finally, and this is the central point for this article we
want to focus on, authorial indexes can be used as devices for dynamic browsing within the documents.

In this paper, we describe the CodeX project (*Consultation et Organisation de Documents Électroniques à parisX*) which aims at studying and evaluating technologies for implementing an XML-based digital library of dissertations. We first introduce and define the term "multiple views" which is the core of our approach and we then present the technical choices we made to develop this user oriented approach. For most users, there are two familiar methods to access the content of the document: the kind of index at the back of the document (often called back-of-book indexes) and the table of content (TOC). In this paper, we focus on *authorial indexes* and we present the CodeX XML-based capabilities which allow the reader to browse the document.

2. Principles

When dealing with large documents, search functions usually do not provide enough information to decide whether or not a given document is relevant. On the other hand, if the user is provided with the entire document (usually in the form of Postscript or PDF documents) he/she is confronted with problems encountered when reading in the online condition.

As it has been pointed out by several studies comparing the reading of paper versus on-line documents in the psychological and ergonomics (Dillon, 1991, 1992), (Hass, 1992) (O'Hara, 1997) (Van Dijk, 1980), documents are not easily readable on a monitor. As a matter of fact, reading in the on-line condition is slow, laborious, and detracted from real conditions of reading. For Landoni (2000) “Table of Contents and Index(es) are essential features. These cannot be simply replaced by search facilities whose complexity makes the reading process more confusing for the user”.

For most of users, the most two familiar access method are those that’s been around the longest: the kind of index at the back of a book and table of content (TOC). There are two general classes of indexes: indexes that are intended for human readers to browse (often called *back-of-book* indexes) and indexes that are intended for use with information retrieval software.

They are some work that addresses the problem of automatically creating an index:

- Turner (1997) presents an empirical evaluation of four algorithms for automatically extracting keywords and keyphrases from documents. The four algorithms are compared using five different collections of documents.
- Woods also used phrase analysis in addition to a large knowledge base to organise terms into a concept hierarchy (1997).
- Evans (2001) present Intellindex which is a software tool that automatically identifies a list of index terms.
- Muñoz (1996) uses an unsupervised learning algorithm to discover two-word keyphrases. The algorithm is based on Adaptive Resonance Theory (ART)
neural networks. Muñoz's algorithm tends to produce a large list of phrases, so it has low precision.

The main feature that distinguishes a back-of-book index from a keyphrase list is length. As Nakagawa (1997) observes, a document is typically assigned keyphrases, but a back-of-book index typically contains index terms. Also, keyphrases are usually intended to cover the whole document, but index terms are intended to cover only a small part of a document. The usability of electronic indexes has been investigated by (Milstead, 1994).

Having an author write, an index is often less expensive. Authors also have a greater understanding of the audience, the vocabulary, and the theory of the material included in the text. In Codex, the authorial index are structured with XML capabilities.

XML offers a possible delivery format for controlled vocabularies. XML can be also used as an alternative to offer a single interface to the different controlled vocabularies and classification systems (Light, 2001). Light (2001) presents some controlled vocabularies encoded in XML. VocML (Vocabulary MarkUp Language) is a DTD under development to support structured representation of a wide range of knowledge organization resources, "including authority files, hierarchical thesauri (including those with polyhierarchies), classification schemes, digital gazetteers, and subject heading lists. Arnold (2001) presents an architecture to place index reference from control vocabularies in XML documents. Luk et al. (2002) provide a survey of current indexing and searching techniques for XML Documents. The indexing techniques described are divided into three groups: flat-file, semistructured, and structured indexing paradigms.

As most electronic thesis present an index (generated automatically by software like Word), Codex create a set of links from the authorial index to the words in the ETD document. Such index is represented in XML. In this approach, the design of the automatic hypertext generation is based on the content-based analysis of the ETDs. This new kind of view does not derive from the logical structure of documents but relies on the textual content. Content browsing applications provide information seekers with access to texts. Structured lists of index terms give a preview of the content of an ETD.

Our feeling is that the benefits of using authorial indexes in a XML Digital Libraries are:

• Less time spent reading: the list of indexes produces a direct access to specific and interrelated passages of the document.
• Less time spent searching: searching with index-terms is a much easier task than doing full-text searching which requires knowledge of queries formulation.
• Better use of an often neglected source of information (user-created index).

3. Technical aspects of the CodeX prototype

During the Word to XML conversion, the position of each user-created index entry in the document logical structure is computed and recorded using XPath expressions. Based on this information, Java servlets running on a Apache server can then build index-based views. Users specify the text they wish to browse; CodeX locates the XML file and associated metadata, determines which portion of the file contains the desired text section, applies appropriate styling rules, and presents the text in HTML form.
Currently, the CodeX system consists of three modules:

- a converter: the converter is a batch tool used to transform our collection of word documents (students' theses) into well-formed XML documents;
- a metadata creation tool: the metadata creation tool is a simple electronic form which is used to generate an TEI-based XML header associated with the documents;
- a visualization module: it uses the concept of "multiple views" to display documents generated by the two other modules in a wide variety of visual forms according to the informational needs of the users. Among the views defined in the first phase of the project (Chaudiron, 2000), there are the following:
  - visualisation of the content table;
  - visualisation of the introduction, conclusion and content table;
  - visualisation of the abstract made by the author;
  - visualisation of the abstract made by the author and the bibliography.

The search and retrieve procedure has the following steps:

- searching metadata database using the web form (see figure1)
- selecting a particular document
- browsing and visualising the content of selected document (see figure 2)

4. Conclusion and future work

The Codex project demonstrates that the XML format can successfully be used to provide electronic access to theses and dissertations. XML provides a standard way for information providers to add custom markup to information-rich documents, in order to publish complex documents in a dynamic way. Codex tools were developed to extract structural and descriptive metadata from these documents and deliver document fragments on demand.

This paper presents also a methodology for indexing collections of ETDs to help users find and browse information more efficiently. Our hypothesis is that browsing ETDs with authorial indexes improves user satisfaction and reduces time of reading. In future, we will compare the three methods for providing index terms (indexers, users, automatic).
Figure 1: Codex search interface

Figure 2: The user browses the structured document entitled “Besoins et usages de résumés” (upper frame) by requesting all occurrences of the user-created index entry “analyse de surface” to be displayed (lower frame).
Notes
http://xml.coverpages.org/vocML.html

References
http://www.infoloom.com/gcaconsns/WEB/TOC/philadelphia99toc.htm”.
(available from http://flood.nhm.ac.uk/mcg/talks/rl/).
Xavier Polanco  
Unit of Research and Innovation, INIST-CNRS, France  

Clusters, Graphs, and Networks for Analysing Internet-Web Supported Communication within a Virtual Community

Abstract: The proposal is to use clusters, graphs and networks as models in order to analyse the Web structure. Clusters, graphs and networks provide knowledge representation and organization. Clusters were generated by co-site analysis. The sample is a set of academic Web sites from the countries belonging to the European Union. These clusters are here revisited from the point of view of graph theory and social network analysis. This is a quantitative and structural analysis. In fact, the Internet is a computer network that connects people and organizations. Thus we may consider it to be a social network. The set of Web academic sites represents an empirical social network, and is viewed as a virtual community. The network structural properties are here analysed applying together cluster analysis, graph theory and social network analysis.

This is a work having taken place in the EICSTES project. EICSTES means European Indicators, Cyberspace, and the Science-Technology-Economy System. It is a research project supported by the Fifth Framework Program of R&D of the European Commission (IST-1999-20350)

1. Introduction

We are concerned with Web structure analysis. Clusters, graphs, and networks procedures will be used to achieved it. The Internet is a computer network that connects people and organizations. As Garton et al. (1997) say "When a computer network connects people or organization, it is a social network." These authors argue the usefulness of a social network approach for the study of computer-mediated communication (CMC). This is our case here. The computer network is the Internet, and people or organizations are represented by a sample of 791 European Union academic Web sites. A community is called virtual community when it is a computer-supported social network in which communication among people is computer-mediated. Our proposal is to analyse the structure properties of the computer-mediated communication (CMC) within a virtual community. In this article, the proposition of using clusters, graphs and networks as models for analysing the Web structure will be examined.

Researchers from either information science or computer science converge to analyse the Web in terms of graph theory (Chakrabarti et al., 1999; Border et al., 2000; Barabasi, 2001), and social network approach (Garton et al., 1997; Chakrabarti, 2000). We first clustered the network data (Polanco et al., 2001) and now we try here to aggregating clusters into a network. With regards to standard social network analysis this is an opposite approach. In social network analysis, the network is the prior given unit of analysis and cluster recognition corresponds to step of partitioning networks into subgroup components. In the case present we start from clusters and use the concepts of graph theory to recognize networks. Another
difference in our approach is that we do not analyse the Web as a graph directly as it is usual in the Web studies.

2. Web Data Sample

The empirical reality is the Web, in our case a sample of 791 academic Web sites from the 15 countries belonging to the European Union, which are grouped into 37 clusters. The problem is the structural analysis of this empirical reality that is a large network in itself. Clusters are our starting blocks of analysis. Now the clusters will be decoded in terms of graph theory and social network analysis.

The Web is a valued directed graph whose nodes correspond to static pages and whose arcs correspond to hyperlinks between these pages. A directed graph consists of a set of nodes, denoted $V$ and a set of arcs, denoted $E$. Each arc is an ordered pair of nodes $(i,j)$ representing a directed connection from $i$ to $j$. The out-degree of a node $i$ is the number of links from $i$ ($i,j_1)\ldots(i,j_k$), and the in-degree is the number of links to $i$ ($j_1,i)\ldots(j_k,i$). From this reality, we have built a representation in which patterns that are hidden in the first reality they become uncovered, and constitute the building block of the analysis.

791 Web sites
5,819,674 Hyperlinks:
• $(i,j$ ($j,i$) = 5,308,204 out-links and in-links (91%)
• $(i,i$ = 511,470 self-links (9%)
12,595,809 pages

Table 1. Web Data Set. These data were collected in January 2001 by M. A. Boudourides and his co-workers at the Computer Technology Institute of Patras, Greece, as part of the project EICSTES.

Let us recall briefly how the sample of 791 academic Web site has been grouped into 37 clusters (see Polanco et al., 2000). The data matrix is a $N$-square matrix noted $D$ where $N$ is equal to the number of sites considered in the data set. The data matrix $D$ recorded the number of hyperlinks between the $N$ sites, in the diagonal the self-hyperlinks denoted $(i,i)$, in the rows the directed hyperlinks denoted $(i,j)$, $(j,i)$ between the $N$ Web sites. From this data matrix we may directly analyse the Web as a directed graph consisting of a set of nodes with directed arcs between pairs of nodes. For the study of the Web in directed graph terms, see (Broder et al., 2000). Though our choice was to built another reality represented by a co-occurrence matrix. In this matrix patterns that are hidden in the data matrix become uncovered, and constitute the building block of the analysis. The hidden pattern is the co-occurrence of a pair of sites in the set of hyperlinks.

3. Co-Site Analysis

The approach that we adopted and we have called co-site analysis consists in recognizing couples of sites. Co-site is defined as the frequency with which two sites are co-associated together in the out-hyperlinks of a set of sites. Co-site analysis agrees with patterns of co-citation (Small, 1973, 1999) and co-word analysis (Callon et al., 1983, 1986). These approaches follow a general co-
occurrence model. The difference is at the level of the object considered to co-
 occur. Thus it is possible to uncover a relationship among a pair of items (authors,
 words, or sites) that do not exist to first and directly approach.

Co-site is a relationship which is established by the hyperlinks of the other
 sites. The co-site frequency of two sites can be determined by computing lists of
 outgoing hyperlinks between the sites in the Web. Each of the two sites is located in
 the set of hyperlinks between sites in a given Web sample, and the number of sites
 at the origin of the hyperlinks defines the frequency of co-occurrence between the
 two sites. A new linking item is simply a new Web site that has hyperlinks with
 both sites. Co-site is the frequency with which two sites are associated together by
 the hyperlinks of the other sites.

In measuring co-site strength, we measure the degree of relationship or
 association between sites as perceived by the population of hyperlinking sites.
 These patterns can be changed over time because of the dependence on the
 hyperlinking sites. Such as vocabulary co-occurrences can change as subject
 domains evolve. The hyperlinking sites are those initiating the hyperlink that
 terminates at the sites receiving the hyperlink. Just as the distinction between citing
 and cited in co-citation analysis. Co-site patterns change as the interest and
 information exchange patterns of the considered field in the Web change.

When two sites are frequently associated by the hyperlinks of the other
 sites, they are also necessarily frequently pointed by the hyperlinks of the other sites
 individually as well. Frequently pointed sites maybe represent the key Web sites in
 a given domain. Thus co-site patterns can be used to map out the relationships
 between these key Web sites. This allows a way of modelling the communication
 structure of a particular set of Web sites. Changes in the co-sites patterns, when
 they are considered over a period of time, may provide clues to understanding the
 mechanism of a Web domain development.

Network of co-sites can be generated for specific sectors of the Web, and
 then submitted to clustering. Clusters of Web co-sites provide a new way to study
 the Web structure. In general a clustering method attempts to reorganize some
 entities into relatively homogeneous groups. Thus clusters represent groups of
 highly similar Web sites. Similarity may be based on the degree of relationship or
 association that exists between entities. We used an association coefficient to
 measure degree of similarity between sites described by binary data, 1 refers to the
 presence of a variable and 0 to its absence. The clustering method that we used was
 a hierarchical agglomerative method following a single linkage rule. From this
 result, a set of clusters, the task is now to produce a network of clusters, a cluster
 network following graph theory.

A cluster consists of five sets of information: [1] a set of components, [2] a
 set of internal associations between pairs of components (or co-sites), [3] a set of
 external associations between pairs of components belonging to different clusters,
 [4] a set of values attached to the associations, and finally [5] a set of sites
 clustered. All the associations or relationships have a strength value. These relations
 are based on the co-occurrence. Each cluster represents itself a valued graph. In the
 next section we deal only with inter-cluster relations in graph-theoretic terms.
4. A Network of Clusters

Following Degenne and Fossé (2001), Wasserman and Faust (1999), we shall revisit mainly the relations between clusters, called inter-cluster relations, according to graph-theoretic concepts. Let each cluster be a node, \( n \), thus \( n \) is equal to 37 nodes. An illustration of this co-site network is shown in Figure 1. Each numbered box in the network diagram represents a cluster of co-sites. The inter-cluster relations are both directional and valued relations. Thus the network of clusters would be represent by a valued directed graph. For simplifying, the graph that is displayed in Figure 1, however, is this that conforms to the following restriction. A binary directed graph, in which the strength values as well as the amount of relations between two clusters are not considered. This restriction forms a natural starting point for modelling a network of clusters. Since it introduces a minimum amount of arbitrary structure whilst still allowing meaningful questions to be asked of the network as whole.

4.1 Directed Graph

The inter-cluster relations are directional relations. A relation is directional if the relation is oriented from one node to another. Directed relations between pairs of nodes are represented as lines in which the directions of the relations are specified by the arrowheads. These oriented lines are called arcs. An arc is an ordered pair of nodes reflecting the direction of the relation between two nodes.

The clusters can be considered as nodes of a directed valued graph. A directed graph, \( G_d(N,L) \) consists of two sets of information: a set of nodes \( N = \{n_1, n_2, \ldots, n_N\} \), and a set of arcs, \( L = \{l_1, l_2, \ldots, l_L\} \). Since each arc is an ordered pair of nodes, there are \( N(N - 1) \) possible arcs in \( L \). Each arc is an ordered pair of distinct nodes, \( l_k = (n_i,n_j) \). The arc \( (n_i,n_j) \) is directed from \( n_i \), the origin or sender node, to \( n_j \), the terminus or receiver node. A node is incident with an arc if the node is in the ordered pair of nodes defining the arc. The nodes \( n_i \) and \( n_j \) are incident with the arc \( l_k = (n_i,n_j) \). Since an arc is an ordered pair of nodes, we can distinguish the first from the second node in the pair, and we must consider if a given node is sender or receiver in the ordered pair defining the arc. Formally, node \( n_i \) is adjacent to node \( n_j \) if \((n_i,n_j) \in L\), and node \( n_j \) is adjacent from node \( n_i \) if \((n_i,n_j) \in L\).

In Figure 1, the clusters are represented as numbered boxes and the arcs are represented as directed arrows. The arc \((n_i,n_j)\) is represented by an arrow from the point representing \( n_i \) to the point representing \( n_j \). For example, if cluster \( i \) has a relation with cluster \( j \) there is an arc originating at \( i \) and terminating at \( j \). If cluster \( j \) returned the tie, there is another arc, this one originating at \( j \) and terminating at \( i \).

In a directed graph, or digraph, three types of relations occur between the \( N(N - 1)/2 \) pairs of nodes: [1] mutual, with both nodes directing relations toward each other, shown by two-headed arrows \((n_i \leftrightarrow n_j)\); [2] asymmetric, in which one node directs a relations toward another that is not reciprocated \((n_i \rightarrow n_j)\); [3] null, in which no relation in either direction exists between a pair of nodes. Figure 1 shows all three types of relations. These patterns can be observed in Figure 1. One might also observe that the graph is far from complete. A graph is said to be complete if all \( N(N-1) \) possible relations between the set of \( N \) nodes are present.
Figure 1. The directed graph representing all the 37 clusters and their interrelations. Each numbered box represents a cluster of co-sites. The directions of the inter-cluster relations are specified by the arrowheads.

4.2 Valued Graph

As mentioned above, the network of clusters consists of valued relations in which the strength of each relation is recorded. In the case of valued relations, valued graphs are the appropriate graph-theoretic representation. A valued graph or a valued directed graph is a graph (or digraph) in which each line (or arc) carries a value. A valued graph, \( G(N,L,V) \), consists of three sets of information: a set of nodes (or vertex, or points), \( N = \{ n_1, n_2, \ldots, n_N \} \), a set of lines (or arcs or edges), \( L = \{ l_1, l_2, \ldots, l_L \} \), and a set of values, \( V = \{ v_1, v_2, \ldots, v_V \} \). Associated with each line (in a graph) or each arc (in a digraph) is a value from the set of real numbers. In our case, the values result from the formula \( E(i,j) = C(i) \times C(j) \), where \( C(i) \) and \( C(j) \) the number of occurrences of sites \( i \) and \( j \) as receivers in the set of the hyperlinks of the other sites.

In a valued graph the relation between node \( n_i \) and node \( n_j \) is identical to the relation between node \( n_j \) and node \( n_i \), \( k = (n_i,n_j) = (n_j,n_i) \), and thus there is only a single value, \( v_k \), for each unordered pair of nodes. It is the case of intra-cluster structure. It is not the case of the relations among the clusters, or inter-cluster relations, which obey an ordered pair of nodes as effects of the clustering method. The order is imposed at the level of clusters by the order of creation of clusters. The relation itself is an undirected valued relation having a single value, \( v_k \), for each pair of nodes each one belongs to two different clusters.

4.3 Valued Directed Graph

The network of clusters should be represented as a valued directed graph in which each cluster is a valued undirected graph. A valued directed graph or a
weighted digraph represents a directional valued relation, such as the amount of links from each Web site to each other Web site. Site $i$ may address a different amount of links to site $j$ than site $j$ address to site $i$. In a valued directed graph, the arc from node $n_i$ to node $n_j$ is not the same as the arc from node $n_j$ to node $n_i$, $l_k = (n_i, n_j) \neq l_m = (n_j, n_i)$, and thus there are two distinct values, one for each possible arc for the ordered pair of nodes. In general, for $l_k = (n_i, n_j)$ and $l_m = (n_j, n_i)$, $v_k$ does not necessarily equal $v_m$.

![Diagram of a digraph](image)

Figure 2. The sub-graphs in which the network may be partitioning. The arcs connecting them have been remote. The removed arcs are the following (see Figure 1): (17→9), (10→6), (10→8), and (14→4), (26→14), (25→16). All the arcs represent single directional relations between the clusters.

The inter-cluster relations among the 37 clusters can be analysed in terms of a valued directed graph. The values that we can consider are the strength values of each relation between a pair of sites, and the out-degree and in-degree values for each cluster. For simplifying, the graph that is displayed in Figure 1, however, is a binary directed graph, in which the strength values as well as the amount of relations between two clusters are not considered. This graph is built from the $N$ square matrix of out-degree and in-degree binary values for each cluster where $N$ is equal to the number of clusters, that is, $N = 37$. In-degrees and out-degrees are useful measurements in particular for our type of networks and relations in which an information exchange occurs.

In a graph, the degree of a node, denoted by $d(n_i)$, is the number of nodes adjacent to it, equivalently, the number of lines incident with it. In a digraph, a node can be either adjacent to, or adjacent from another node, depending on the direction of the arc. It is interesting to consider these cases separately. Since one quantifies the tendency of nodes to be senders; the other quantifies the tendency to receive. A node with degree equal to 0 is called an isolate.
• **Odal out-degree.** The out-degree of a node, \( d_o(n_i) \), is the number of nodes adjacent from \( n_i \). The out-degree of node \( n_i \) is equal to the number of arcs of the form \( l_k = (n_o,n_j) \), for all \( l_k \in L \), and all \( n_j \in N \). The out-degree value measures the number of arcs originating with any node \( n_i \).

• **Nodal in-degree.** The in-degree of a node, \( d_i(n_i) \), is the number of nodes that are adjacent to \( n_i \). The in-degree of node \( n_i \) is equal to the number of arcs of the form \( l_k = (n_i,n_o) \), for all \( l_k \in L \), and all \( n_i \in N \). The in-degree value measures the number of arcs terminating at any \( n_i \).

In our example of inter-clusters relationships, clusters with high out-degree can be recognized as *senders*, and clusters with high in-degree as heavy *receivers* in the exchange of information. There are also isolated clusters. Note these clusters are each one a set of co-sites. This is a way to discover the isolated subset of sites within the network. *Isolate* means here to be close inside of a cluster without connections with any other cluster. Also pairs of clusters are distinguished; they are only related together without relations with the other clusters. Thus they are together isolated from the others in the network.

In terms of the in-degrees and out-degrees of the nodes in a directed graph, we can distinguish four different kinds of nodes (Wasserman and Faust, 1999). [1] The node is an isolate; [2] the node only has arcs originating from it; [3] the node only has arcs terminating at it; [4] the node has arcs both to and from it. All these cases can be observed in our example as shown in Figure 1.

5. **A Network of Networks**

Finally, we may consider the overall structure of the network took as a whole. From this point of view, three sub-networks can be recognized in Figure 1. These three sub-networks are shown in a separate manner in Figure 2. The arcs connecting them have been remote. This is a way of analysing the whole network. Furthermore, the information that each cluster represents allows to know the academic sites that are together in each sub-network. We can also to known the cities and countries in which they are located. The example is chosen mainly to illustrate how the whole network can be interpreted in a real situation. We can also to expect to observe changes in the network patterns, when they are considered over a period of time. This approach may provide clues to understanding the mechanism of a Web domain development.

In this framework, many different indices can be computed from matrices to measure structural characteristics for both individual actors and entire networks. This issue is a important subject that is out of the scope of this article. We then stop here.

6. **Conclusions**

Let us recall that our research proposal deals with using *clusters*, *graphs* and *networks* as models for analysing the Web structure. This is in progress. In this article, we have limited to use graph theoretic concepts for analysing clusters of co-sites. The issue that remains to be considered in the framework of pattern
recognition and exploratory data analysis is the clustering methods based in graph theory. Just as this is treated in Hubert, (1974), Dubes and Jain, (1980), Theodoridis and Koutroumbas, (1999).

It appears that an interpretation of the significance of the clusters of co-sites must rely on the notion of similarity, and on the association or co-occurrence of contents. Co-site clusters maybe correspond to significant intellectual connections within the Web field in consideration. This suggests extending the co-sites analyse into Web content analysis.

Another area for the application of co-site analysis is in the study of the structure of science in the Web. We are thinking in the co-site analysis restated as we began to do here in the framework of graph theory and social network analysis. The pattern of relations among key R&D sites establishes a structure for the scientific specialty which may then be observed to change through time. Through the study of these changing structures, co-site analysis becomes a tool for analysing and monitoring the development of scientific fields on the Web, and for assessing the degree of interrelationship among specialties in the Web context, as well as co-citation and co-word analysis in the context of bibliographic databases since a long time in the study of science.

References
An IRS Based on Multi-Granular Linguistic Information

Abstract: An information retrieval system (IRS) based on fuzzy multi-granular linguistic information is proposed. The system has an evaluation method to process multi-granular linguistic information, in such a way that the inputs to the IRS are represented in a different linguistic domain than the outputs. The system accepts Boolean queries whose terms are weighted by means of the ordinal linguistic values represented by the linguistic variable "Importance" assessed on a label set $S$. The system evaluates the weighted queries according to a threshold semantic and obtains the linguistic retrieval status values (RSV) of documents represented by a linguistic variable "Relevance" expressed in a different label set $S'$. The advantage of this linguistic IRS with respect to others is that the use of the multi-granular linguistic information facilitates and improves the IRS-user interaction.

1. Introduction

Information Retrieval (IR) is a research field referred to the storage and retrieval of textual information (Korfhage, 1997; Salton, 1989). IR systems (IRSs) carry out two main activities: i) to store documents by computing index term weights and ii) to retrieve documents by matching user queries and documents. An important question in the IRSs is how to facilitate the IRS-user interaction.

The use of linguistic variables (Zadeh, 1975) to represent the input and output information in the retrieval process of IRSs improves considerably the IRS-user interaction (Bordogna and Pasi, 1993) (Kraft et al., 1994) (Herrera-Viedma, 1999, 2001). Usually, the most linguistic IRSs assume that users provide their information needs by means of Boolean queries whose terms are weighted by linguistic values represented by the linguistic variable "Importance" assessed on a label set $S$. IRSs evaluate the linguistic weighted queries and provide the linguistic RSVs of documents represented by the linguistic variable "Relevance" assessed on the same label set $S$. The drawback is that the use of the same label set to express the inputs and outputs of linguistic IRSs diminishes the communication capability in the IRS-user interaction. Furthermore, the above linguistic variables represent different concepts, and thus, it seems necessary to use different linguistic expression domains to model them.

In this paper, we present a linguistic IRS that manages multi-granular linguistic information using an ordinal fuzzy linguistic approach (Herrera et al., 1996) (Herrera-Viedma, 2001). The weighted Boolean queries and the RSVs of documents are assessed on label sets with different granularity and/or semantics. The query terms are weighted according to a threshold semantic. The Boolean operators AND and OR are modeled by means of the linguistic aggregation operator, LOWA operator (Herrera et al., 1996). The LOWA operator is an and-or
operator, and this property allows us to introduce a soft computing in the evaluation of queries. The retrieved documents are arranged in linguistic relevance classes, which are identified by ordinal linguistic terms.

To do so, the paper is structured as follows. Section 2 is devoted to introduce the ordinal fuzzy linguistic approach, the concept of multi-granular linguistic information and the LOWA operator. Then, the IRS based on multi-granular linguistic information is presented in Section 3. Finally, several conclusions are pointed out in Section 4.

2. The Ordinal Fuzzy Linguistic Approach

The ordinal fuzzy linguistic approach is an approximate technique appropriate to deal with qualitative aspects of problems (Herrera et al., 1996). An ordinal fuzzy linguistic approach is defined by considering a finite and totally ordered label set \( S = \{s_i, i \in H = 0,...,T\} \) in the usual sense and with odd cardinality (7 or 9 labels). The mid term representing an assessment of "approximately 0.5" and the rest of the terms being placed symmetrically around it. The semantic of the linguistic term set is established from the ordered structure of the term set by considering that each linguistic term for the pair \((s_i,s_T-i)\) is equally informative. For each label \(s_i\) is given a fuzzy number defined on the \([0,1]\) interval, which is described by a linear trapezoidal membership function represented by the 4-tuple \((a_i, b_i, a_i, \beta_i)\) (the first two parameters indicate the interval in which the membership value is 1.0; the third and fourth parameters indicate the left and right widths of the distribution). Furthermore, we require the following properties:

1. The set is ordered: \( s_i \geq s_j \) if \( i \geq j \).
2. There is the negation operator: \( \text{Neg}(s_i) = s_{T-i} \), with \( j = T - i \).
3. Maximization operator: \( \text{MAX}(s_i, s_j) = s_i \) if \( s_i \geq s_j \).
4. Minimization operator: \( \text{MIN}(s_i, s_j) = s_i \) if \( s_j \leq s_i \).

2.1. On Multi-Granular Linguistic Information

In any linguistic approach, an important parameter to determine is the granularity of uncertainty, i.e., the cardinality of the linguistic term set \( S \) used to express the information. The cardinality of \( S \) must be small enough so as not to impose useless precision on the users, and it must be rich enough in order to allow a discrimination of the assessments in a limited number of degrees.

On the other hand, according to the uncertainty degree that a user qualifying a phenomenon has on it, the linguistic term set chosen to provide his knowledge will have more or less terms. When different users have different uncertainty degrees on the phenomenon, then several linguistic term sets with a different granularity of uncertainty are necessary. Then, we need tools of management of multi-granular linguistic information to model these situations. Different proposals can be found in (Delgado et al., 1998) (Herrera et al., 2000).

In Delgado et al. (1998) we characterize some transformation functions between the linguistic and numerical expression domains using the concept of the characteristic values associated to a label.
Let us consider that for each label \( s_i \) we know a set of characteristic values, \( CV_i = \{ C_1^i, C_2^i, \ldots, C_z^i \} \), which are crisp values that summarize the information given by \( s_i \), i.e., they support its meaning. We shall assume that \( C_j^i \in \text{Supp}(s_i) = \{ r \in \mathbb{R} | \mu_i(r) > 0 \} \). Without loss of generality, we can define a set of functions \( CF = \{ f_j, j = 1, \ldots, z \} \), in such a way that each function \( f_j \) associates a characteristic value to each label \( s_i \), i.e., \( f_j : F(\mathbb{R}) \to \mathbb{R}, f_j(s_i) = C_j^i \), being \( F(\mathbb{R}) \) the set of fuzzy numbers defined on \( \mathbb{R} \) that we can use to characterize the semantic of the labels. Some examples of this function type are:

- **The defuzzification method of gravity center** (Cordon et al., 1997):
  \[
  f_j(s_i) = \left[ (b_i + \beta_i)^2 + (b_i)^2 - (a_i)^2 - (a_i + \alpha_i)^2 - (b_i + \beta_i)b_i - a_i(a_i + \alpha_i) \right] / 3(2b_i + \beta_i - 2a_i + \alpha_i), \]
  and \( f_j(s_i) = a_i + \alpha_i \) if \( \beta_i + \alpha_i = 0 \).

- **The maximum value** (Delgado et al., 1998b):
  \[
  f_j(s_i) = \max \{ v | \mu_i(v) = \text{Sup} \{ \mu_i(t), \forall t \} \}. \]

**Definition 1.** The linguistic-numerical transformation function, \( A^N \), for any label \( s_i \) is defined according to the following expression: \( A^N : S \to [0,1], A^N(s_i) = g(f_1(s_i), f_2(s_i), \ldots, f_z(s_i)) \), being \( g \) any aggregation operator verifying: \( \min \{ v_1, v_2, \ldots, v_z \} \leq g(v_1, v_2, \ldots, v_z) \leq \max \{ v_1, v_2, \ldots, v_z \} \).

Therefore, \( A^N \) obtains the real value of a label by means of the aggregation of its respective characteristic values. An example of \( g \) can be the mean function.

**Definition 2.** The numerical-linguistic transformation function, \( A^L \), for any numerical value \( r \in [0,1] \) is defined according to the following expression: \( A^L : [0,1] \to S, A^L(r) = s_i \), being \( s_i \) a label verifying: \( h(r, s_i) = \min \{ h(r, s_p) \mid \forall s_p \in S \} \), with

In this paper, we use the above transformation functions to define a tool for processing multi-granular linguistic information in the retrieval process of IRS.

### 2.2. The LOWA Operator

The Linguistic Ordered Weighted Averaging (LOWA) is an aggregation operator of ordinal linguistic values based on symbolic computation (Herrera et al., 1996). It acts by direct computation on the labels only taking into account the order of linguistic assessments without considering the associated membership functions.

**Definition 3.** Let \( A = \{ a_1, \ldots, a_m \} \) be a set of labels to be aggregated, then the LOWA operator, \( \Theta \), is defined as \( \Theta(a_1, \ldots, a_m) = \sum_{j=1}^{m} \theta_j \sum_{k=1}^{m} \sum_{\sigma(j) < \sigma(k)} w_{jk} \Theta \sum_{\sigma(j) = \sigma(k)} w_{jk} \Theta \sum_{\sigma(j) > \sigma(k)} w_{jk} \Theta \), where \( \sum_{j=1}^{m} \sum_{k=1}^{m} \sum_{\sigma(j) < \sigma(k)} w_{jk} \Theta \sum_{\sigma(j) = \sigma(k)} w_{jk} \Theta \sum_{\sigma(j) > \sigma(k)} w_{jk} \Theta = A \), is a vector associated to \( A \), such that, \( \tau = \sigma(A) \) = \( \{ a_{\sigma(1)}, \ldots, a_{\sigma(m)} \} \), where \( a_{\sigma(i)} \leq a_{\sigma(l)} \forall i \leq l \), with \( \sigma \) being a permutation over the set of labels \( A \). \( C^m \) is the convex combination operator of \( m \) labels and if \( m=2 \), then it is defined as \( C^2 \{ w_h b_h, i = 1, 2 \} = w_i \Theta \sigma j \Theta (1 - w_i) \Theta \sum_{j=1}^{m} w_{jh} \Theta b_j, \Theta \sum_{j=1}^{m} w_{jh} \Theta b_j, \Theta \sum_{j=1}^{m} w_{jh} \Theta b_j, \Theta \), such that \( k = \min \{ T, i + \text{round}(w_j (j - i)) \} \), \( j, s_j, s_i \in S, (j \geq i) \), being "round" the usual round operation, and \( b_j = s_j, b_j = s_i \). If \( w_j = 1 \) and \( w_i = 0 \) with \( i \neq j \forall i \), then \( C^m \{ w_h b_h, i = 1, \ldots, m \} = b_j \).
The LOWA operator is an "or-and" operator (Herrera et al., 1996). This property allows that the LOWA operator carries out a soft computing in the modelling of MAX and MIN linguistic operators. We use this good characteristic in our linguistic IRS to evaluate the Boolean queries. In order to classify OWA operators in regard to their localisation between and and or, Yager (Yager, 1988) introduced a measure of orness, associated with any vector W as follows

\[ \text{orness}(W) = \frac{1}{m-1} \sum_{k=1}^{m} (m-k)w_k. \]

Fixed a W, then the nearer an OWA operator is to an or, the closer its orness measure is to one; while the nearer it is to an and, the closer is to zero. Generally, an OWA operator with much on nonzero weights near the top will be an or-like operator (orness \( \geq 0.5 \)), and when much of the weights are nonzero near the bottom, the OWA operator will be an and-like.

3. The IRS Based on Multi-Granular Linguistic Information

In this section we present an IRS that accepts linguistic weighted Boolean queries, supports multi-granular linguistic information and models the Boolean operators in a flexible way.

We assume that the documents \( D=\{d_1,...,d_m\} \) are represented by means of index terms \( T=\{t_1,...,t_n\} \). Each term has associated an index term weight \( F \) which describes the subject content of the documents. \( F: D \times T \rightarrow [0,1] \) is a numerical indexing function that maps a given document \( d_j \) and a given index term \( t_i \) to a numeric weight between 0 and 1. \( F(d_j, t_i) \) is a numerical weight that represents the degree of significance of \( d_j \) in \( t_i \). \( F(d_j, t_i)=0 \) implies that the document \( d_j \), is not at all about the concept(s) represented by index term \( t_i \), \( F(d_j, t_i)=1 \) implies that the document \( d_j \) is perfectly represented by the concept(s) indicated by \( t_i \), and \( F(d_j, t_i) \in (0,1) \) represents the different intermediate significance degrees.

3.1. The Linguistic Weighted Boolean Queries

In this IRS each query is expressed as a combination of the weighted index terms which are connected by the logical operators AND (\( \wedge \)), OR (\( \vee \)), and NOT (\( \neg \)) and weighted with ordinal linguistic terms represented by the linguistic variable "Importance" assessed on a label set \( S \). Thus, as was done in (Herrera-Viedma, 2001), we assume a set of ordinal linguistic terms \( S \) to express the linguistic weights.

In this context, a query is any legitimate Boolean expression whose atomic components (atoms) are 2-tuplas \( <t_i, c> \) belonging to the set, \( T \times S \); \( t_i \in T \) (set of index terms), \( c_i \) is a label of the linguistic variable "Importance", modelling a threshold semantic. Therefore, the set \( Q \) of the legitimate linguistic weighted Boolean queries is defined by the following syntactic rules:

1. \( \forall q =< t_i, c_i> \in T \times S \rightarrow q \in Q \).
2. \( \forall q, p \in Q \rightarrow q \wedge p \in Q. \)
3. \( \forall q, p \in Q \rightarrow q \vee p \in Q. \)
4. \( \forall q \in Q \rightarrow \neg (q) \in Q. \)
5. All legitimate linguistic weighted Boolean queries \( q \in Q \) are only those obtained by applying rules 1-4.
3.2. Evaluation Procedure of User Queries

In this subsection, we present how IRS evaluates a user query in a multi-granular linguistic framework, that is, assuming that the values RSV assigned to the documents are represented by means of the linguistic variable “Relevance” which is assessed on a label set $S' \neq S$. To define the evaluation procedure, previously we have to establish the semantics associated to the weights of user queries.

Particularly, we assume that the weights of query terms are associated to a symmetrical threshold semantics (Herrera-Viedma, 2001). This semantics considers that a user can search for documents with a minimally acceptable presence of one term in their representations as in or documents with a maximally acceptable absence of one term in their representations. Then, when a user asks for documents in which the concept(s) represented by a term $t_i$ is (are) with the value High Importance, the user would not reject a document with a $F$ value greater than High; on the contrary, when a user asks for documents in which the concept(s) represented by a term $t_i$ is (are) with the value Low Importance, the user would not reject a document with a $F$ value less than Low. In practice, given a request $<t_i, c_i>$, this means that the linguistic query weights that imply the presence of a term in a document $c_i \geq s_{72}$ (e.g. High, Very High) it must be treated differently to the linguistic query weights that imply the absence of one term in a document $c_i \geq s_{72}$ (e.g. Low, Very Low). Then, if $c_i \geq s_{72}$ the request $<t_i, w>$, is synonymous with the request $<t_i, c_i>$, which expresses the fact that the desired documents are those having $F$ values as high as possible; and if $c_i < s_{72}$ is synonymous with the request $<t_i, c_i>$, which expresses the fact that the desired documents are those having $F$ values as low as possible.

Then, the evaluation procedure evaluates a linguistic weighted Boolean query in three steps:

1.- Making uniform the information using the transformation functions given in Subsection 2.4. This implies that the numerical index term weights of the documents and the linguistic weights of queries must be expressed in the domain of the linguistic variable “Relevance”.

2.- The documents are evaluated according to their relevance only to atoms of the query applying the symmetrical threshold semantic.

3.- The documents are evaluated according to their relevance to Boolean combinations of atomic components, and so on, working in a bottom-up fashion until the whole query is processed.

Then, the evaluation procedure is modelled by a linguistic matching function $E^*: Q \times D \rightarrow S'$ that, for a given $q \in Q$ yields for each $d_j \in D$ an ordinal linguistic value $RSV_j = E^*(q,d_j) \in S'$. $E^*$ is defined recursively applying the following rules:

1.- $E^*(q,d_j) = g*(\Lambda^L (F(d_j,t_i)), \Lambda^L (\Lambda^N (c_i))), \forall q=<t_i, c_i>, \forall j, \Lambda^L : [0,1] \rightarrow S'$, $\Lambda^N : S \rightarrow [0,1]$, and $g*: S' \times S' \rightarrow S'$ is the linguistic matching function that models the symmetrical threshold semantics (Herrera-Viedma, 2001):

$K$ is a sensitivity parameter defined to control the importance of the closeness between $\Lambda^L (F)$ and $w_i$ in the final result. The greater the value of $K$, the smaller the importance of the value of distance. $K = 1$ means that the symmetrical threshold semantic is not used.
\[ g'(s_b, s_a) = \]

<table>
<thead>
<tr>
<th>( s_0 )</th>
<th>if ( s_b \geq s_T/2 ) ( \land ) ( s_a = s_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s_1 )</td>
<td>if ( s_b \geq s_T/2 ) ( \land ) ( s_0 &lt; s_a &lt; s_b )</td>
</tr>
<tr>
<td>( s_2 )</td>
<td>if ( s_b \geq s_T/2 ) ( \land ) ( s_b \leq s_a &lt; s_T )</td>
</tr>
<tr>
<td>( s_T )</td>
<td>if ( s_b \geq s_T/2 ) ( \land ) ( s_a = s_T )</td>
</tr>
<tr>
<td>( S_n )</td>
<td>if ( s_0 \geq s_T/2 ) ( \land ) ( s_a = s_0 )</td>
</tr>
<tr>
<td>( S_i )</td>
<td>if ( s_0 \geq s_T/2 ) ( \land ) ( s_0 &lt; s_a &lt; s_b )</td>
</tr>
</tbody>
</table>

\( i_1 = \max\{0, \text{round}\left(\frac{b-((b-a)/K)}{K}\right)\} \)

\( i_2 = \min\{T, \text{round}\left(\frac{b+((b-a)/K)}{K}\right)\} \)

\( K \in \{2, 3, 4, \ldots, b\} \).

2.- On the negated queries, \( \rightarrow q \). We assume that the evaluation procedure can only deal with negated atoms \( \neg t_i, c_i \). This may be easily achieved applying the De Morgan’s laws on any query. Then, we define the evaluation of document \( d_j \) for a negated weighted atom \( \neg t_i, c_i \) from the negation of index term weight \( F(d_j, t_i) \): \( E^*(q, d_j) = g'(\neg g(A^k(F(d_j, t_i))), A^k(c_i)) \).

3.- \( E^*(\land k \leq M_{2} q_k, d_j) = \Phi( E^*(q_1, d_j), \ldots E^*(q_M, d_j)) \), using a weighting vector \( W \) in such a way that \( \text{orness}(W) < 0.5 \).

4.- \( E^*(\lor k \leq M_{2} q_k, d_j) = \Phi( E^*(q_1, d_j), \ldots E^*(q_M, d_j)) \), using a weighting vector \( W \) in such a way that \( \text{orness}(W) \geq 0.5 \).

4. Conclusions

We have presented a linguistic IRS that supports the use of different label sets to express system inputs (user queries) and outputs (RSVs). In such a way, we have improved the IRS-user interaction.

In the future, we shall study how to improve the performance of IRSs based on multi-granular linguistic information by means of other different tools of processing of information.

References


Pedro Cuesta, Alma M. Gómez and Francisco J. Rodríguez
Department of Computer Science, University of Vigo, Ourense, Spain

Using Agents for Information Retrieval

Abstract: In this work the use of information agents in an information retrieval environment is introduced. Initially, a brief revision of some basic theoretical aspects is presented. The contributions in this aspect are: a classification for non-cooperative information agents, based on tasks that the agent has to perform, and the identification of main functionalities which must be provided in cooperative information agents systems. In addition, from a practical point of view, MASIR, a cooperative agents prototype, which provides uniform access through Internet to heterogeneous and dispersed documental databases is explained.

1. Introduction

Nowadays, the advances in Information Technology and Communications and particularly the Internet revolution, allows accessing to a huge amount of information. As the Web size increases, finding out the suitable information becomes a more difficult task. Other problems related are the changes in data (pages appear and/or disappear or are frequently modified), redundancy (duplicated contents) or quality (incorrect or erroneous information). In addition, the problem complexity grows due to the distributed and heterogeneous nature of information. Data sources differ in interface, format, structure, and language, and are distributed among a great number of connected computers.

Information retrieval (IR) techniques try to solve some of the previously introduced problems. These methods provide a way of locating relevant documents based on user requirements (Baeza-Yates, 1999). Besides, advanced systems which help in efficient and unified access to information are needed. This has led to the appearance of new flexible and powerful tools and sophisticated algorithms (currently implemented in search engines) which help users in the process of retrieving relevant information.

In an environment the described before, intelligent agents constitute a key technology for the future of Internet. Agents may assist the user in finding useful, relevant information, managing and overcoming the difficulties associated with "information overload" (Klusch, 1999). They may also inform the user that new relevant data have been published, negotiate a sell or buy of products, participate in electronic auctions, etc. Agents may carry out these tasks independently or working in a coordinated way with other agents.

In this work the system MASIR, which is an example of cooperating agents is introduced. The paper is organised as follows. Section 2 introduce the theoretical aspects of intelligent information agents; while in part 3 concepts about non-cooperative agents are addressed. In section 4, cooperative agents, which is the key technology in the system is exposed in depth. Section 5 focuses in the description of MASIR prototype. Finally, in point 6 conclusions and future work are presented.
2. Intelligent Information Agents

Agents represent an important advance in abstraction level in the process of software development. They may be used for the design and implementation of complex (distributed) software systems (Jennings, 1999). Actually there is no universal definition of Agent concept, but a widely used one states that an intelligent agent is a computer system capable of flexible, autonomous action in some environment; in this definition, flexible means that agent is reactive; pro-active and social. (Wooldridge, 1995).

Intelligent Information Agents can be defined as computational software systems which access to multiple heterogeneous and geographically distributed information sources in order to help users in the process of searching relevant information (Klusch, 1999). These agents provide transparent access to many different information sources, simplifying the problem of managing a large amount of data. This implies that agents must interpret the user request, analyse and translate it in the correct way for each source, retrieve the desired information, integrate the results and give them back to the user. All this process must be done preferably in a just-in-time fashion.

Information Agents may be classified according to different fashion. Depending on their ability to co-operate with each other in the execution of tasks, agents may be non cooperative or cooperative. This aspect distinguishes two kinds of systems, the ones which are formed by a single agent and systems composed by a society of interacting agents. Any of the previous types may be classified as rational, which are utilitarian in an economic sense, adaptive which are able to adapt themselves to changes in the environment or mobile which travel autonomously trough the Internet (Klusch, 1999).

3. Non Cooperative Information Agents

Klush classification is based in the characteristics of the agent internal behaviour. Nevertheless, we thought it is possible to introduce a different classification taking into account the tasks that the agent has to accomplish. This classification is closer to final user point of view, and it is similar to the ones used in some Internet agents repositories (Botspot, Agentland, ...). According to this criterion, several kinds of non-cooperative information agents can be distinguished:

- **Search agents**, help the user to retrieve information from heterogeneous and distributed sources. They provide a fast and simple way to obtain relevant information. Examples of search agents are Bullseye (http://www.intelliseek.com), CiteSeer (http://www.researchindex.com) and Copernic (http://www.copernic.com/).

- **Monitor agents** control the changes in different information sources (for instance, changes in web structures, updated news in a newspaper...) or the appearance of new sources related with a particular subject in order to warn the user. Usually, the notification of changes is sent to user by e-mail. Examples of monitor agents are Mind-It (http://minder.netmind.com/) and Informant (http://informant.dartmouth.edu/)

- **Filter agents** reduce the incoming information keeping only the relevant data from the user point of view. The preferences of the user are stored in his personal profile. Examples of filter agents are InfoScan
Browser agents help the user in navigation through the Web. It is usual to have browser agents integrated into the navigator. They highlight dead links, emphasised keywords in the web page, anticipate relevant links, etc. Interquick (http://interquick.deerfield.com/) and Letizia (Lieberman, 1995) are examples of browser agents.

Agents for electronic commerce offer commercial services, in order to save time and money. They are able to recommend a certain product, to compare prices among different electronic shops, etc. Some examples are MySimon (http://www.mysimon.com/) and Pricerunner (http://www.pricerunner.com/).

4. Cooperative Information Agents

Many of current investigation related to intelligent agents has focused on individual agent capabilities and structure. Nevertheless, in order to solve complex problems these agents must cooperate with other agents in a heterogeneous environment. A multiagent system (MAS) is a set of agents which communicate among them and cooperate in task execution (Sycara, 1998). This interaction allows solving problems that are beyond the individual capabilities of each individual agent. Interaction in MAS differs from other computing paradigms since it takes place at knowledge level using a high-level agent communication language like KQML, FIPA-ACL (Labrou, 1998). Basically, the advantages of an approach based on cooperative agents are: simplicity, flexibility, robustness, scalability and the integration of existing legacy systems.

In this cooperative agents environment two main functionalities must be provided: a mechanism for linking the different agents and a way of solving the heterogeneity of managed information.

4.1. Intermediation

Cooperation among information agents can be achieved using different interaction techniques. Taking into account the role played, three different types of agents can be distinguished (Klusch, 2001).

1. **Provider agents** offer their capabilities to users and other agents.
2. **Requester agents** use information and services offer by provider agents.
3. **Middle agents** mediate for a correct communication among providers and requesters.

The process of mediation done by middle agents is based on the following steps: (1) Provider agents advertise their capabilities to one or more middle agents, describing the service they provide. (2) Middle agents are responsible for storing all these advertisement. (3) A requester agent asks for locating and connecting to provider agents which offer a desired service. (4) The middle agent using the stored advertisements returns the result.

Depending on the kind of result returned, two types of middle agents are distinguished:

1. **Matchmaker agent**, the result is an ordered list of provider agents which offer the requested service. Once this result is received by requester agents, it is the responsible for contacting the provider agent, negotiating and performing the transaction.
2. Broker agent, in contrast with matchmaker, performs the complete transaction. This means that there is no direct communication between provider and requester agent, because all the operations go through the broker agent. The main tasks of these agents are contacting appropriate provider agents, negotiating, performing and controlling the transaction and return the results of service to the requester agent.

Given the fact that different types of middle agents provide different performance trade-offs what types of middle-agents are appropriate depends on the application (Klusch, 2001).

4.2. Ontologies

One of the main problems for cooperation in an agents society is the semantic heterogeneity of information that the agents must manage. Semantic heterogeneity considers the content of an information item and its “intended” meaning. In order to achieve this semantic heterogeneity, the meaning of the interchanged information has to be understood across the agents society. An ontology can be defined as an explicit specification of a conceptualisation (Gruber, 1994); that is, a representation (with a set of concepts and the relationships among them) of an abstract and simplified view of the world.

In agents societies, ontologies can be used to describe the semantics of the requests and service descriptions and to make explicit the content of the different information sources. They also reduce conceptual and terminological conflicts providing a unified framework.

The middle-agent overcomes the semantic heterogeneity by means of a knowledge-based process which relays on using ontologies. That is, the use of ontologies enable shared understanding among different agents with different aims and different viewpoints of the global system.

4.3. Examples of Cooperative Multiagent Systems

RETSINA is an open multi-agent system, which performs an information retrieval and integration directed by goals, in support of a variety of decision-making tasks (Sycara, 2001). The process of mediation in RETSINA basically relies on service matchmaking. The specification of capability and service is based on an Agent Capability Description Language (ACDL) called LARKS (Language for Advertisement and Request for Knowledge Sharing). Application domain knowledge in agent advertisements and requests can be currently specified as local ontologies written in a specific concept language ITL or by using WordNet.

InfoSleuth is an agent-based system that can be configured to perform many different information management activities in a distributed environment (Bayardo, 1997). It is composed of seven kinds of software agents which, all together, provide a number of complex query services which involve solving ontology-based queries over the dynamically changing resources. InfoSleuth architecture uses a process of matchmaking, that is, the broker agent maintain a knowledge base with information about all the agents in the system and their capabilities and uses this knowledge to match required services with the provider agent.

5. MASIR

Multiagent System for Information Retrieval (MASIR) is a prototype based in cooperative information agents. The system allows uniform access trough Internet to several heterogeneous documental databases. The prototype allows, by
means of an unique interface, to pose a query, send it to several databases, get the answers from the databases and finally integrate the results and present them to user. The scalability of the system permits the integration of new sources, with independence of their data model, structure, query language, DBMS or location.

The global architecture of the system is presented in Figure 1. The system is integrated by four kinds of agents: interface, wrapper, broker, and ontology.

- Each user interacts with the system through its own interface agent. It is responsible for capturing the query, translate it to the suitable FIPA-ACL sentence (Request) in order to send it to broker agent and facilitate the user navigation through the obtained results. The interface agent, making use of the information supplied by ontology agent, generates dynamically, in order to make the query and to present the results, a user domain specific interface.

- The ontology agent contains an abstraction of the databases schemata, linking each concept to the databases where it is relevant. The ontology agent also provides to interface agents and broker the domain terms for the query and the semantic relationships. A more detailed description can be found in (Cuesta, 2001).

- The broker agent is the system core and acts as an intermediary among user and the various documental databases. It takes the request received from interface agent, composes subrequests, and distributes them among wrapper agents (using FIPA-ACL Contract-Net). In order to achieve this objective, it asks the ontology agent what the databases implied in the request are. It also synchronises answers to query, manages detected errors, integrates the answers received from several wrapper agents and sends them back to interface agent.

- The wrapper agent knows the specific features of its related source. It translates the subrequest received from broker agent in order to access its
associated documental database and obtain the required data. Each request may imply, for example, the execution of SQL or OQL sentences or the utilisation of text retrieval techniques. In addition, the wrapper agent contains all the details of the semantic knowledge for that source. When a new source is added, a wrapper agent will be created associated with it; meanwhile the ontology agent will update its knowledge base.

The comparison of MASIR with the systems presented in point 4.3, can be made taking into account the intermediation model and the use of ontologies. From the first point of view, both RETSINA and InfoSleuth are based on a matchmaking process. This implies a direct interaction between Requester agents and selected Provider agents, which is performed independently from the matchmaker. This avoids, for example, data transmission bottlenecks or single point of failure at the matchmaker but increases communication overhead among agents. MASIR, in contrast uses the brokering model. The main reason is avoiding communication overhead due to the necessity of solving source heterogeneity through the use of ontology agent.

From an ontologically point of view MASIR has a higher level of description than RETSINA or InfoSleuth because MASIR maintains a global agent ontology. This agent provides an abstraction of the information sources schemas, linking each concept to the specific source where it is relevant. RETSINA use only local ontologies to specify agent advertisements and requests instead of using them to integrate the information sources. On the other hand, InfoSleuth has a set of specific (local) domain ontologies and does not provide way to integrate the local ontologies.

6. Conclusions and future work

In this work, the advantages of using agents for information retrieval has been presented. Besides, a new classification for non-cooperative information agents has been proposed; this taxonomy is based on tasks that the agent has to perform. In the construction of cooperative agents systems, two important aspects must be taken into account: the model for intermediation among the different agents and how ontologies are used to solve the heterogeneity of information.

In contrast with other systems, MASIR uses a middle agent with brokering intermediation and a global ontology agent for solving the semantic conflicts. The system is able to grow; when a new source is added, an associated wrapper agent is created and the ontology is actualised incorporating the source new semantics.

As future work, many aspects will be studied. One of the key challenges is related with the communicational aspect of MASIR; at the moment, the interaction protocols are being refined. At ontological level, the development of mechanism for facilitating ontology creation and modification is being studied. In addition, the adaptation of MASIR to “FIPAOntology-Server” is currently under consideration, in order to allow the integration of external ontologies. Another current work area is the user of adaptative interface agents with capability of learning from the user behaviour and adapting its knowledge model to facilitate customisation and suitable task delegation.
References
Knowledge Organization in a Multilingual System for the Personalization of Digital News Services: How to Integrate Knowledge

Abstract: In this paper we are concerned with the type of services that send periodic news selections to subscribers of a digital newspaper by means of electronic mail. The aims are to study the influence of categorisation in information retrieval and in digital newspapers, different models to solve problems of bilingualism in digital information services and to analyse the evaluation in information filtering and personalisation in information agents. Hermes* is a multilingual system for the personalisation of news services which allows integration and categorisation of information in two languages. In order to customise information for each user, Hermes provides the means for representing a user interests homogeneously across the operating languages of the system. A simple system is applied to train automatically a dynamic news item classifier for both languages, by taking the Yahoo set of categories as reference framework and using the web pages classified under them as training collection. Traditional evaluation methods have been applied and their shortcomings for the present endeavour have been noted.

1. Introduction

The recent boom in the popularity of the Internet has resulted in a rapid expansion of the range of information services available to the common user. One such service is that of systems offering to send users a selection of the daily news by e-mail. New ways of understanding information services and information systems are arising. In this paper we are concerned with the type of services that send periodic news selections to subscribers of a digital newspaper by means of electronic mail.

The task of managing the volume of information that the advent of Internet has thrust into our hands faces two significant challenges. The first challenge is posed by the ever present globalisation, which demands a capability for dealing with information in several languages in a homogeneous manner. The second challenge is a much older problem but made severe by the sheer volume of material currently in circulation: how to classify documents with a minimum of effort in order to provide subsets of the whole to which a user interested in a particular topic can address himself without having to shift through the complete set. Once a system attempts to face both challenges at the same time, the problem grows. The main
question to be faced is how to improve on a rough and ready initial classification of
documents under language heading (documents in English and documents in
Spanish) to achieve a classification by topic independently of the language
employed. This may present additional problems of granularity of the classification,
due to the fact that fine grained classifications in different languages soon lose any
semblance of similarity that coarse grained classifications may have had. At a
certain level Spanish categories for news items will branch off into a bullfighting
section, whereas the English equivalent may branch off to cricket or baseball. This
is not entirely a linguistic problem and is probably more related with cultural issues,
but the problem remains and must be addressed.

2. Resources for Multilingual Information Access

It is very important for multilingual search to take into account both the
growth of information services and monolingual digital libraries and the need for
tools with multilingual capacity for information retrieval and extraction (Abaitua,
2000). An effective global information transfer faces up to the challenge posed by
the large number of national languages in use. Language differences may become a
barrier to information circulation in the world, among persons and among
organizations. The access to foreign-language information can be facilitated by
multilingual glossaries, thesauri and classifications (they can provide multilingual
pointers to the subject matter of documents), and translations (Lancaster, 1992).

The use of bilingual corpora is very interesting in the development of
applications - as in terminology, automatic translation, and information multilingual
search -, especially over the Internet. There are different kinds of multilingual
corpora: corpora of texts in different languages to implement quantitative or
statistical studies; comparable corpora, consisting of texts in a language and
translations of similar documents in the same language; and parallel corpora, the
same collection of texts in more than a language. Explicit correspondence
relationships should be made between segments of each language, by means of
grammatical categories.

Asghar and Revie (2000) provide an interesting discussion of the role of
thesauri and classifications in Internet: the growth of information in the worldwide
Web and the migration of information resources to the new context demand a better
and consistent subject identification; thesauri and classifications collaborate on
description of information resources, avoiding problems associated with quality of
information retrieved in the Web; thesauri and classifications improve the rapid and
ey easy access to the information in the Web.

Approaches to the construction of a new multilingual thesauri are: usual
construction of a thesaurus, seeking equivalencies among terms collected (with
different results among languages), without direct references to terms or structures
of an existing thesaurus; translation of a monolingual thesaurus; conciliation and
adaptation of existing thesauri in two or more languages. In truth, multilingual
access to document collections is crucial. Besides, the co-operation improves the
instruments connected with the information retrieval and the access to the
information, in order to facilitate human and automatic indexation and to create
links among related institutions (Lancaster, 1992; Clavel-Merrin, 1999).

According to Aitchison and Gilchrist (1990), after verifying the suitability of the
project, terms and categories of the thesaurus are translated with their equivalents.
Documents in the source language are analyzed to assign them to categories (classification) or assigning different terms to each document in order to represent and to facilitate its retrieval (thesaurus). The last step is the formulation of the query in another language. By means of an automatic system, the user can search for terms with the equivalent terms in the original language as query.

3. Multilingual Information Access in Hermes

Hermes is a system that applies existing techniques from the field of text classification, text categorization (Sebastiani, 1999) and information retrieval (Salton, 1989), besides user modelling (Amato & Straccia, 1999), to the selection of items, from different newspapers in different languages (Spanish and English), relevant for a user. Each user can create a profile in his language with his preferences and receive daily the news items that interest him from the different newspapers (Diaz et al., 2000).

A user accesses the information server and registers for the service. The user selects his language and different data about his preferences (email address, days of the week to receive news, maximum number of items per message) and interests. These interests are: the sections of the newspapers, an alternative system of classification (first level of categories from Yahoo), and terms chosen by the user as interesting.

The system manages two models per user, one per language, and applies each model to the news in the same language. The categories of Yahoo are language independent because there is a hierarchy in each language with the same first level categories. The terms are translated to from one language to the other.

The message received by the user contains: the name of the user, the date, and a list of news items ranked according to the user information interests and respecting the maximum number of items per message defined. Each news item is presented with the source, the author, the title, a short summary adapted to the user (Acero et al., 2001), the relevance, and a link to the news item in the digital newspaper. At the end of the message appear the interests of the user as features in his profile in order to allow the user to check the true relevance of the received news.

Finally, the system allows relevance feedback (Nakashima and Nakamura, 1997). The user can vote about the news in a positive, in a negative or in an indifferent way. This information is captured by the system in another interest for the user, the feedback terms that will be used in the next selection of news item.

4. Multilingual Text Classification in Hermes

Hermes uses three different systems for classifying information: one is the static classification of news items into sections provided by the newspaper domain, a second one is provided by a dynamic classification of the news items carried out automatically in terms of the categories used in the Yahoo directory, and a third one may be provided by the user as a custom-tailored category defined by a set of keywords and which is also automatically applied to the news items. The final classification is obtained by combining these sources through a weighted formula, according to a set of weights specified in the user model during configuration. These systems should ideally be as orthogonal as possible, in order to present truly
different classifications of the domain. This is not the case altogether, but the overlap is not excessively significant.

4.1 The Choice of Categories

The categories of Yahoo were chosen as a reference framework in the first approximation for various reasons generally related with the overall efficiency of the process. On one hand, they come associated with distinct sets of classified documents in different languages (those classified under the English and the Spanish versions of Yahoo). These sets of documents were easily accessible in electronic form and could be used to train the automatic classifier to be employed. On the other hand, they are a set of categories specifically designed to facilitate search through a heterogeneous collection of documents, such as is found in the web. It was hoped that the differences between the set of news items in your run-of-the-mill daily edition and the collection of documents available in the web would ensure that this second set of categories add information to the existing one in terms of newspaper sections.

Various problems come associated with this choice. The automatic classifier is trained with documents corresponding to a domain other than the domain of application. The branching structure of the hypertext documents classified under each category implies that it is not always clear what page is an actual good example (possibly only leaves of the resulting hypertext trees should be used, cropping those intermediate pages which simply substructure a given category into subcategories but hold no relevant content themselves), and this introduces a degree of noise in the classification system. The effect of these problems in the evaluated results has been noted, and they are currently being explored in search of an optimized solution.

4.2 Dealing with More than One Language

In Hermes each user builds a model defining his preferences over categories and keywords for a single language, and the system generates a model in the other language automatically. Information about newspaper sections is not generated in this way because it is language dependent. This is a clear instance of equivalence problems between languages, made even more acute by the fact that each newspaper may have its own set of sections, even if working in the same language. The technique employed for generating models in a different language is based on the translation of the keywords defined by the user. The use of Yahoo categories, together with the assumption that Yahoo categories across different languages match, simplifies the process. Once the models for the two languages have been built, the news items for each language are processed with respect to the corresponding version of the model. Each of the language specific classification processes is independent of the other.

The final classification is carried out by combining the three different sources of classification through the weighted formula. Where automatic classification is required, it is achieved by calculating the one-to-one similarity between news items and the representation of the categories using the cosine formula of the Vector Space Model (Salton, 1989).

The representation of each category is obtained by training with different documents associated to that category (Sebastiani, 1999). A possible solution to the problems outlined above concerning the disparity of domains resulting for this particular choice of set of categories would be to train the system with a manually classified set of real news items, but classified under the Yahoo system. This would
represent an important volume of work and would lose the advantages of having a dynamically updated set of sample documents for the chosen categories, with matching representation in different languages. Alternative solutions would be to combine both types of documents in training, or to perform co-training (Blum & Mitchell, 1998) on the representation of the categories, using the daily set of correctly classified news items. Either solution would gather together the advantages of both approaches.

5. Evaluation of Multilingual Information Systems

Evaluation of these new instruments requires: a reflection about categorisation, a validation of traditional evaluation measures within the new field of Internet, the consideration of the knowledge acquired during evaluation of search engines, and a close study of the working principles and the required evaluation according to the particular properties and conditions of the service under consideration.

Although there are various procedures for the evaluation of information systems, the emergence of the particular combination of challenges, objectives and techniques involved in personalised news services gives rise to additional issues that need to be addressed during system evaluation. On one hand, these systems have to ensure that the tools they provide for the user to specify his interest in information items of a particular type are sound according to traditional information retrieval measurements. On the other hand, they face a competitive market where different methods of specifying user interest are continuously competing for the user's eye, so any particular technique being employed must prove its worth in terms of user satisfaction. The following aspects must be covered in a thorough evaluation:

a) categorisation, filtering, personalisation.

b) user response

c) the vision that users develop of the system

d) user profiles

e) values of recall and precision for all the users on several specific days

In order to achieve all these aims, explicit evaluations provided by the users are harvested for feedback on system response-time, ease of use, system efficiency, and conceptual and physical presentation. This information is compiled on the basis of a closed questionnaire with specific questions on the relevant main topics. The user is asked to evaluate aspects such as category overlap, category validity, relevance of a document for the assigned category, or quality of the overall category scheme.

Additionally, a manual analysis of news items and user models logged by the system for a set of chosen days is carried out in terms of classic information retrieval measurements, which provide quantitative values for system efficiency.

The experience of evaluating system performance and user satisfaction for different personalised news services (Diaz et al., 2000) has proven the importance of the nature of the information in this tasks, the relative merits of the three most popular methods of specifying information interests (sections, categories, and key words) with respect to this particular set of tasks, and the risks of careless application of recall and precision measures in systems such as these where different methods of specifying interests are combined (Diaz et al., 2001).
An initial evaluation of a prototype of our system has given good feelings about the performance. This evaluation has been developed using a working pattern adapted to a monolingual version of the system used in previous experiments. This pattern includes several aspects as interface evaluation, newspaper sections, categories, summaries, bilingual capacity and user estimated recall and precision.

In general, users found the system suitable. They are satisfied with the different aspects of the user model, they estimate that the translation of the keywords is sometimes less than adequate but they value in a positive way the possibility to receive news in different languages.

We have yet to perform a more complete evaluation with a larger number of users and the relations between the different features that appear in our system must be studied in greater detail. For instance, how the multilinguality and the user modeling affect the traditional way of evaluating information retrieval systems, i.e. recall and precision measures.

6. Conclusions

This system can be a powerful tool in a multilingual context. In a globalized environment information services may take a principal role in overcoming linguistic and knowledge barriers, and contributing to the interrelation and even integration of cultures, economies and societies. In truth, this integration depends on the efficiency of the system. The construction of this crucial instrument for the Information Society requires an evaluation that takes into account the user, the impact of automatic categorisation and user modelling, as well as the problems derived from the use of more than one language. Nonetheless, this tool will work in an integrating manner, from a cultural and knowledge perspective, whenever the contents that it helps to retrieve are specifically structured for this purpose - for instance, by respecting the differences between the different cultures, and supporting the common ground.

Acknowledgements

This work has been partially supported by the Spanish Ministry of Science and Technology (Ref. 2000/020)

References


Knowledge Representation and Organization of Gender Studies on the Internet: Towards Integration

Abstract: This study analyzes the models of representation and organization of knowledge surrounding Gender. In addition to assessing their visibility, we analyze the thematic models and the conceptual treatment of Gender in search engine directories with wide coverage in different parts of the world. Eight search engines were selected, two of them international (Yahoo, Google), one an international affiliate (Yahoo España), and five local ones (three from the Mercosur, and one each from Africa and Asia). The research was done on two levels: a) knowledge representation through the quantitative and qualitative analysis of the terms of Gender; b) knowledge organization, through the structural and semantic-contextual analysis of the search engines. The results express a clear terminological and structural supremacy of the international search engines, as well a lack of correspondence overall among the terminologies, relative visibility, and diffusion of matters of Gender, leading to considerable difficulties in achieving consistent access to specific information.

1. Introduction

Gender studies get steady attention from many different disciplines. This is reflected in publishing trends—not only in conventional forms, but also in a great deal of Internet documents. Gender is a particularly sensitive and complex subject for knowledge organization and representation because in it, scientific and disciplinary aspects are tied up with others that pursue historical, social and cultural claims. Exploring Gender-related themes entails special difficulties because their development depends on conceptions, preferences and prejudices that are socially established, conditioned or refuted concerning the situation of women and men. For this reason, the representation of the topics as conceptual structures to be used by persons from different cultures can be very problematic, and calls for retrieval models based on culturally integrated knowledge.

The definition of Gender presents an additional difficulty, as it is a multi- and inter-disciplinary expression that has not yet firmly established its scope and epistemological borders. For some authors, Gender is “the set of symbolic, social, economic, legal, political and cultural attributes assigned to persons in accordance with their sex” (Lagarde, 1995). Others affirm that gender refers to “the set of social characteristics attributed to a person according to his or her sex. Biological sex is not the same as Gender, which is the identity assumed by or acquired by that sex” (Ferro Calabrese, 1998). A more exhaustive development of the evolution of
Gender studies is described by other authors (De Torres and Muñoz, 2000, p. 13ss.).

The main objective of the present study is to identify the means of Organization and Representation of Gender Knowledge that are used in the category tables of Internet search engines, in standardized language, to guide or facilitate access to and the retrieval of Web information.

Our review of literature on this subject reaffirms the need for studies of this sort, as no antecedents could be found. Moreover, no other areas dealing with knowledge organization on the Internet have systematically explored Gender from this perspective.

Topics of relatively recent interest always present specific obstacles for Knowledge Organization. In the case of Gender:

Different societies, nations and communities face Gender themes in variety of ways, in the light of cultural, religious and social conceptions or models, or the world view of that particular culture or subculture.

The documentary languages available on this subject are scarce, and the divisions of knowledge that they make vary considerably. There are no clear lines around the thematic space covered; that is, the existence of diverse interpretations of the interior structure of the theme reveal a conceptual inconsistency overall.

There are still not many libraries specializing in gender studies, and their development is restricted, above all in Less Developed Countries. No research to date contrasts the conceptual structures and terminologies with the documentary reality of the theme.

It may also be said that this study ties in with the study of minorities, to which knowledge organization has thus far given little attention. In this particular case, women as central objects of attention in gender studies are considered a cultural and social minority, because just as other groups and social levels, they suffer situations of scarce social visibility and forms of relegation and discrimination based on sex.

2. Objectives and principles

The following objectives were defined for this research study:

To determine the current state of visibility, and analyze the way the theme Gender is dealt with on the Internet.

- To compare the models established by search engines with wide coverage of the population in different parts of the world, considering the Anglo-Saxon, Spanish, Latin American, African and Asian perspectives.
- To analyze whether the search engines adopt a user perspective when they organize their category tables.
- To contribute to the integration of knowledge on this topic.

The following formal principles were followed in the analysis: Descriptive standpoint. In that Gender is a theme laden with interpretations that lean heavily on the political or ideological realm, we agreed to apply a descriptive analysis of the linguistic and terminological treatment of the term by search engines.
Use of source language. The linguistic formulations of the source were respected, though in some cases contextual clarifications of the context were introduced (indicated in parentheses).

Correlation between terminology and disciplinary development. We departed from the premise that the terms used by search engines could reflect the state of development of gender studies in a given region of the world, expressing the implicit conceptions, prejudices and existing views on the theme.

Respect for variety and integration of variety. Cultural diversity was taken into account, incorporating the different social patterns in which communities express themselves. The approach taken was such that it might contribute to cultural integration in the areas surrounding Gender.

3. Methodology

The research was carried out on two levels: a) knowledge representation, verified through the analysis of the terms used by search engines to identify each label in the category tables; and b) knowledge organization, assessed through the structural and semantic-contextual analysis of the search engine directories.

As a point of reference for terminology in the subject area, we used four specific thesauri produced in different regions: Europe (Information Centre and Archives for the Women’s Movement, 1998), Spain (Instituto de la Mujer, 1999), Catalonia (Sebastiá y Salat, 1988) and Brazil (Bruschini, Ardaillon & Unbehaum, 1998).

Bibliotechnological and terminological criteria applied for data analysis included:

Compilation of terminology pertaining to or associated with Gender proposed in the category tables, and contrasted for validation with the terminology proposed by these thesauri, in order to face the different tasks involving linguistic analysis.

Weighting of terms, to place each under one of the three following headings:

- Terms specific to gender, whose relevance to the theme is clear and unequivocal (for example, Feminism, Women’s Studies, Masculinity).
- Terms on themes related or akin to Gender, and terms traditionally associated with one sex (e.g. Fashion, Gynecologic Disorders). Many terms of this type are linked by the search engines to sites destined to specific users, as women (for example, Beauty).
- Disciplines including aspects of gender, under which specific concepts of gender appear (e.g. Education, Sexology, Society).

Formal expressions were excluded (e.g. History, Institutions, Publications, Periodicals; Associations).

The core terms, that is, those most pertinent to Gender for the purposes of this study, were considered to be:
a) those relative to ideologies and movements interpreting Gender in the context of human rights and social equality;
b) those referring to matters that affect men and women because of their Gender, related for instance with their sexuality, marital status, social, political, economic and cultural circumstances;
c) terms pertaining to the Family and Family Relations, whenever they have a specific connection with gender;
d) terms that appear under categories such as Sexuality and Sexology.

Terms bearing a more distant relation with the nucleus were considered to be akin terms, such as:
- terms referring to the Family and Family Relations that have no specific link with gender (for example, Domestic Violence, Surrogate Mothers, At Home Dads);
- terms traditionally associated with women, but not exclusively woman-oriented in the modern organization of knowledge (Fashion, Beauty, Models);

In the areas of Health and Psychology, we restricted selection to include only those terms strongly relevant to Gender. Yet in general, we included as terms of a subject area those concerning both sexes indiscriminately (e.g. Infertility), or those referring to one gender with respect to the other (e.g. Wives).

The word term was preferred as the general denominator because the usual meaning of term (an expression made up of one or more words; as opposed to descriptor, key word, or heading) was more practical for the search engine category tables.

Whenever possible, the top term established for a Gender-related subject was identified.

A total of 8 search engines were selected, on the basis of geographical distribution and presentation, to ensure satisfactory coverage of different countries, cultural conceptions, religious ideas, and behavioral patterns. At the same time, we chose some familiar local search engines that would have a wide public potential in specific regions of the world (India, South Africa, the MERCOSUR countries, Spain and the US). Expressly excluded from the analysis for the sake of consistency were gender-specific search engines (such as Femina). The search engines finally chosen were:

Two international search engines from the English-language realm: Google and Yahoo.

A Spanish subsidiary of an international search engine: Yahoo España.


The following aspects were analyzed:

From the standpoint of Knowledge Organization, the general structural design of the search engines was studied, as was their possible influence on the visibility of the theme Gender. Special attention was given to the structure of the organization of Gender-related knowledge: the disciplinary location of Gender and its degree of autonomy or subordination; the identification of the strongest disciplinary associations; the concentration or diffusion of the Gender-related subjects; direct or indirect access to the sites; and the degree of development, the contextual treatment of terminology and the conceptions underlying the associations established.

From the standpoint of Knowledge Representation, we considered the quantitative dimension of the terms, their classification and distribution within the three areas indicated earlier (Gender-specific, akin themes, less directly related disciplines); the identification of the top terms for Gender and their variants; the existence of notes of scope or definitions of a term somewhere within the directory;
the existence of two or more locations for a term, in separate tables or on different levels, and the means of marking them; the disciplinary localization of each Gender term; the indication of the number of sites under each term; and finally, the qualitative analysis of the terminology used and the qualitative analysis of the presentation of categories, identifying conceptual inconsistencies and the particularities of a focus.

For further study, the terms extracted from the directory were comparatively analyzed as to: their frequency or occurrence in the search engines studied; the level of subdivision in which they appear; the analysis of multiple appearances; the synonyms detected; and the correlation of occurrences with the thesauri taken as reference.

4. Results

From the standpoint of Knowledge Organization, the relative visibility and the inconsistencies of handling Gender-related topics are manifest in the following ways:

a) In no case is gender treated as a discipline; when it appears, it is subordinated to main categories such as Culture & Society, Society, Social Science(s), Family, Family & Relationship, Sexology and Health. Other Gender-relevant aspects or subjects appear non-systematically in categories such as Art & Humanities, People, Psychology, Sexology and Sports.

b) Six search engines offer autonomous category tables of their own for Gender, but they are always subordinated to others. Only Cadé Brasil allows direct access to the category of Gender.

c) Though certain search engines do establish associations among the categories of a directory using hypertextual links, the identification of these links is not standardized. However, when two categories that have an associative relationship are presented in a hierarchical fashion, the true nature of the relation is covered up.

Categories on table 2 show the disciplinary dependencies of the Gender terms in each directory. Society, Social Science(s), Health and Sexology have the most representative associations. The diffusion of Gender terms is greater in Google (14 disciplines and subdisciplines), whereas Cadé Brasil appears to have the most homogenous structure (21 core terms and 8 akin terms in 5 disciplines).

From the standpoint of knowledge representation, only the international search engines and their affiliates show a significant representation of the theme. According to table 1, out of 123 core categories, Yahoo includes 78, Google 65 and Yahoo España 27. The local search engines show deficient representation, ranging between 0 and 21 core terms, distributed as follows: Cadé 21, Uruguay Total 7, Directorio Argentino 2, India 4, South Africa 0.

The international search engines feature labels that are excessively repeated and poorly structured. In their quantitative dimension, our results underline: a) the use of 123 terms with 204 occurrences to represent specific Gender themes; and another 32, with 54 occurrences, representing akin themes or those less directly associated with Gender, data which may be useful as a parameter for the elaboration and evaluation of specific documentary languages; b) an overwhelming supremacy of the international search engines and their affiliates insofar as the exhaustiveness of coverage, as mentioned earlier: the three international search engines studied
gave a sum total of 170 of the 204 occurrences of Gender terms (83.33%); in fact, only 4 of the 123 core terms and 5 of the 32 akin terms were not registered by any of the three; c) interestingly, the search engine that provides the broadest terminological coverage (Yahoo) only uses 63% of the gender labels identified (78/123); d) a lack of terminological consistency, as 62% of the core terms appear only once in the directories, and just over 16% occur three or more times; e) a poor terminological correspondence between the two strictly international search engines (Yahoo and Google), who share just 32 out of 123 specific terms (26%) and only 2 of the 32 akin terms (6.25%); and finally, f) Yahoo international was found to coincide with its Spanish subsidiary on just 60% of the specific terms and none of the akin terms. It is also very noteworthy that core terms covered by Yahoo España are limited to only 34.6% of the coverage by Yahoo international (27/78).

Five top terms were identified: Women (3 occurrences), Woman (2), Gender, Women's Studies and Social Interaction (1 occurrence each. The use of the word Gender in labelling is characteristic of the international search engines, though it is used by the Uruguayan search engine as well. The local search engines consistently prefer the term Woman, which might indirectly indicate the stage of development of this subject in the different regions of the world.

From the qualitative point of view we may observe: a) only 2 search engines offer notes on the scope or definition to help the user get conceptually situated; b) only the international search engines and the Spanish affiliate present two or more locations for a single term, whereas in the others Gender terms are associated with only one category or discipline; c) there is an emphasis on the representation of aspects of Gender tied to sexuality (especially with homosexuality), family life or personal relationships; d) standardization is not applied to similar terms used by the search engines (for example, Gay & Bisexual; Gay, Lesbian & Bisexual; Gay & Lesbian(s); Gay, Lesbian & Bisexual Studies; Lesbian, Gays & Bisexual(s) express more or less the same ideas); e) there is no standard marking system for identifying related categories or occurrences of one same category in another part of the directory; and f) 7 of the 8 search engines allow free category searches.

5. Conclusions

Our research demonstrates that there are great difficulties in representing knowledge about a social topic such as Gender on the Internet, if we bear in mind that users belong to different cultures and societies, and that it is crucial for the directories to make use of an international terminology that is minimally standardized and that, at the same time, respects local particularities. A thematic retrieval offering culturally integrated knowledge should be promoted. Notwithstanding, we are facing a situation in which at times these considerations are not incorporated in the knowledge managed by search engines.

The general characteristics of organization and representation of knowledge about Gender may evidence the state of evolution of gender studies in a given culture, and could stand as a measure of cultural evolution and social development—a category is only justified if there are a good number of sites on one theme, and these, in turn, express an interest in the supply and demand of specific information. On the other hand, the choice of top terms by a search engine could also be indicating the degree of development of the theme, as the absence or presence of
themes such as Gender Studies or Women’s Studies may reflect a point in the
evolution of the disciplinary development in a certain geographic region.
There is a significant disintegration and lack of development of Gender as a theme
in the local search engines, all operating from Less Developed Countries. Thus, the
South African search engine presents Gender sites vaguely organized in wider
categories under the label “Social Interaction.” Only the Brazilian directory offers a
somewhat developed conceptual structure.

It is clearly shown that the international search engines and their affiliates
offer a much more detailed and comprehensive treatment of Gender-related themes,
and communicate the most advanced knowledge on the subject.

The choice and location of the Gender terms in a directory vary depending
on the levels of discourse of the search engines which, in turn, are influenced by
their objectives and the perceived target users. This comes to confirm to what extent
subjective aspects (intentionality, objectives, etc. of search engines) influence the
representation and organization of knowledge.

Finally, it is advisable that the designers of directories and search engines take
greater care to make their respective conceptual structures and terminological
representations compatible for the different cultures served.

In short, the results presented here confirm that the theme Gender on the
Internet does not have adequate visibility. Its models of thematic organization,
furthermore, are not necessarily representative of cultural particularities or of the
discourse and disciplinary development, but rather evidence current inconsistencies
and a need for greater integration. As occurs with documentary languages, the
Internet Search engines prove that we do not yet have ready access to directories
with consistent and uniform structures, able to realistically represent Gender
studies.

References
Rica, 1998
Instituto de la Mujer (España): Centro de Documentación (1999). Tesauro “Mujer”.
International Information Centre and Archives for the Women’s Movement (1998).
Lagarde, Marcela (1995) Género y desarrollo de la teoría feminista: enfoque de género. La
Sebastiá i Salat, Montserrat (1988), Thésaurus d’història social de la dona. Barcelona:
Torres Ramírez, Isabel de and Ana María Muñoz Muñoz (2000), Fuentes de información

Web addresses
Ananzi de Sudáfrica www.ananzi.co.za Visited January, 2002
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Abortion</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Abstinence</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Acquaintance Rape</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Anarcha Feminism</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Androgen insensitivity syndrome</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Aphrodisiacs</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Bears</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Birth Control</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Bisexual</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Butch-femme</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Celibacy</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Circumcision</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Cross Dressing</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Divorce(s)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Divorced Fathers</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Drag Kings</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Drag Queens</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Ecofeminism</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Educational gender equity</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>E-Romance</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Erótica</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Extramarital affairs</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Fathering</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Fathers</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Female to Male</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Female's Genital Mutilation</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Feminism</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Feminist pedagogy</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Feminist theology</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Fetishes</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Gay, Lesbian &amp; Bisexual</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Gay, Lesbian &amp; Bisexual Studies</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Gay Men</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Gay &amp; Bisexual</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Gay &amp; Lesbians</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Gender</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Gender Assignement</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Gender Equity</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Gender Law</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Gender Neutral Language</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Gender Studies</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Gender &amp; Sexuality</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Category</td>
<td>X Count</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genderqueer</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genital Mutiliation</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Girls</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grr(r)ls</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heterosexual steem</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homophobia</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homosexuals &amp; Bisexuals</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Husbands</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incest</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infertility</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intersexed</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intersexuality</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lesbian</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lesbian, Gays &amp; Bisexual(s)</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lesbian, Gay &amp; Bisexual Studies</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lesbian &amp; Bisexual</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Love Dating</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lovers</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male Rape</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male to Female</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marital Rape</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marriage</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Masturbation</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matrimonial</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men’s Health</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men’s Movement</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men’s Rights</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men’s Studies</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mothering</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mothers</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parenting</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Politics of Sexuality</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polyamory</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polygamy</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pornography</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pride Celebrations</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pro-Feminist Men</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prostitution</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rape</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reproduction</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reproduction &amp; Sexuality</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Riot Grrls</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Romance</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sacred Sexuality</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Safer Sex</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Term</td>
<td>Count</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>-------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex Crimes</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex Education</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex Therapy</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex Work</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sexs</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sexual Abuse</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sexual Addiction</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sexual Crimes</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sexual Disorders</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sexual Harassment</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sexual Psychology</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sexuality</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sexuality &amp; Relationships</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sexuality &amp; Sex</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sexually Transmitted Diseases (STD’s)</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single Mothers</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single Fathers</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transsexuals</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transgendered</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Venereal Diseases</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wedding &amp; Marriage</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weddings</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Widowers</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Widows</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wives</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women of Color</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women's Health</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women's History</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women's Magazines</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women's Resources</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women's Rights</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women's Studies</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Working Mothers</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**TERMS: 123 78 27 65 21 7 2 4 0 204**

Table 1. Core vocabulary in search engines directories
**DISCIPLINES**

<table>
<thead>
<tr>
<th>DISCIPLINES</th>
<th>Y.</th>
<th>Y. E.</th>
<th>G.</th>
<th>C.</th>
<th>U.T.</th>
<th>D. A.</th>
<th>I.</th>
<th>A.</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts &amp; Humanities</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Conditions &amp; Diseases</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Culture &amp; Groups</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Education</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Health</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>4</td>
</tr>
<tr>
<td>Health &amp; Beauty</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>2</td>
</tr>
<tr>
<td>Health &amp; Wellness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>1</td>
</tr>
<tr>
<td>Obstetrics &amp; Gynecology</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Organizations</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>3</td>
</tr>
<tr>
<td>Psychology</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Relationship(s)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Sexology</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Shopping</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Social Organizations</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>2</td>
</tr>
<tr>
<td>Social Science(s)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Society</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Society &amp; Culture</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Society &amp; Religion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>1</td>
</tr>
<tr>
<td>Sports</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Subcultures</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Supports</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Disciplines :</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>46</td>
</tr>
</tbody>
</table>

Table 2. Disciplines and subdisciplines related to gender in search engines directories

**Abreviations**

- Y: Yahoo
- Y. E.: Yahoo España
- G.: Google
- C.: Cadé
- U.T.: Uruguay Total
- D.A.: Directorio Argentino
- I.: 123 India
- A.: Ananzi de South Africa
Abstract: The existence of huge amounts of information available in information systems and networks worldwide imposes the creation of adequate tools able to efficiently organize it and allow its retrieval across geographical, linguistic and cultural boundaries. An indexing language covering all areas of knowledge and converting the language-independent structure of a classification system like the Universal Decimal Classification into a thesaurus structure in more than one language seems to be a solution. Among the key attributes of the indexing language thus obtained we can mention: consistency in indexing, control on terms, user-friendliness. The paper presents the great potential in information retrieval of the combined retrieval method by means of a case study.

1. Introduction

Among the consequences of the rapid development of the global information society a major one is the existence of huge amounts of information stored in information systems and networks across geographical, linguistic and cultural boundaries. The need was imposed to create tools and technologies able to efficiently organize and allow retrieval of information in this universal context. Information professionals had to cope not only with the multitude of knowledge organisation and representation systems but also with the multitude of languages the available information is stored in order to provide the users with effective information retrieval tools.

For this purpose a real language industry has been developed, theoreticians and researchers making considerable efforts to find feasible solutions to problems of multilingual access by way of natural language processing and machine translation methodologies. Such corporate efforts belong to the CoBRA+ working group for multilingual access to subjects (MACS) or to the cross-language information retrieval (CLIR) tracks of the Text Retrieval Conferences that annually report the progress made in multilingual information access and retrieval. The encouraging results they have obtained so far are still confined to discipline/domain restrictions and most of their achievements are based on language pairs rather than multiple language combinations.

The multilingual access facilities offered by combining two different types of information languages in order to improve information retrieval make the substance of this report. The approach going to be described has two component parts:

- mapping a traditional classification system onto an interdisciplinary multilingual thesaurus
- making information retrieval possible by means of either UDC codes or UDC-based descriptors in any of the languages included in the project.
2. The Universal Decimal Classification and its potential to switch between languages

Along with the remarkable efforts made in the 1970's to prove the benefits of using the UDC as an intermediate or switching language, some other efforts were concentrated in the direction of building alphabetical indexes to the UDC tables in order to facilitate browsing and searching. An example of such an index is one Lorphèvre (1973) made to the French International Medium Edition of the UDC. According to what the author argues in the introduction, this index includes all the expressions in the tables taken in alphabetical order. Yet, he goes on, the absence of some words do not imply that there is no way to classify such subjects but they can be expressed by combinations of numbers according to the UDC grammar rules.

Example:

\[
\begin{array}{ll}
\text{Développement des bibliothèques} & 021 \\
\text{des plantes} & 581.14 \\
\text{équipement photographique} & 771.4 \\
\text{ontogénétique} & 577.95 \\
\text{organique, physiologie} & 612.64 \\
\text{organisation générale} & 65.016 \\
\text{photographie} & 77.023 \\
\text{urbanisme} & 711.12 \\
\text{zoophysiole} & 591.16 \\
\end{array}
\]

The comparison Lorphèvre makes in the introduction between his index and a thesaurus (or "trésor", as he calls it) is what we are concerned with. The first difference is a rather formal one, of succession of the two sections of the indexing tools, i.e. the classifications start with the systematic tables going from general to particular and they are completed by an alphabetical table, while the thesauri go the other way around: the alphabetical part comes first and it is followed by the systematic part. Another difference is relating to the way the coordination of terms is made: in the case of the UDC precoordination is the intrinsic feature of its own structure and additionally there are also "see" and "see also" references provided for this purpose, whereas in the thesaurus the coordination is indicated by the codes used in the alphabetical part. In the end the author announces his intention to transform the alphabetical indexes of the UDC into "trésors", structuring them differently in the first place, and then coordinating the terms. This way there will be no difference, he says, between the two methods, the last word belonging to the UDC "qui se joue des langues qui traduisent ses notations chiffrees".

Once again here, the advantage of numeric systems over alphabetical systems is underlined. Without saying it straightforwardly, the main advantage lies in the lack of ambiguity of the numeric notations as is the case with the problematic use of the natural language words that translate them.

3. Description of the case study

The great potential in information retrieval of the combined retrieval method will be demonstrated by a case study. An overview of the main issues covered by our project is following:
3.1. Structure of the database

A database has been built out of bibliographic records taken from a real-life online catalogue. The multilingual document collection thus constituted has subjects represented initially by UDC notations in all the bibliographic records and manually assigned descriptors in Romanian in some of them. To these bibliographic records the UDC Master Reference File (MRF) and a shortened version of this containing only the upper subdivisions of the UDC classes were added. The latter section of the database is meant to provide context to the captions. All in all, the database has 159553 records. A number of 97922 are bibliographic records with UDC notations, of which 40627 have manually assigned descriptors and 84110 have UDC-derived ones (see Table 1).

<table>
<thead>
<tr>
<th>Number of records in the bibliographic database</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of records in the Master Reference File (MRF)</td>
<td>61457</td>
<td>38.52%</td>
</tr>
<tr>
<td>Number of records in the short MRF</td>
<td>174</td>
<td>0.11%</td>
</tr>
<tr>
<td>Total number of records</td>
<td>159553</td>
<td>100%</td>
</tr>
<tr>
<td>Number of bibliographic records indexed with manually assigned descriptors</td>
<td>40627</td>
<td>25.46%</td>
</tr>
<tr>
<td>Number of bibliographic records indexed with automatically assigned multilingual descriptors</td>
<td>84110</td>
<td>52.72%</td>
</tr>
</tbody>
</table>

Table 1. Configuration of the experimental database

To serve our purposes of integrating a multilingual thesaurus into a bibliographic database in order to facilitate multilingual access, an indexing tool was created from a pre-established list of descriptors based on an abridged version of the UDC. The list of 1254 Romanian descriptors was developed into a trilingual thesaurus in Romanian, English and French that eventually had well over 3,000 terms (descriptors and non-descriptors) in each contributing language and a total of 4827 terms.

3.2. Automatic assignment of multilingual descriptors

Several programs have been created to perform each of the following activities in the experimental database such as:

- integrate the multilingual thesaurus terms in the MRF by matching the UDC-derived descriptors with the UDC numbers in the file;
- merge the resulting product with the bibliographic database to which successive other programs were previously applied in order to decompose the UDC notations given in the classified catalogue into their separate parts and add the corresponding text to them.

The bibliographic record in Figure 1 illustrates the additional access points to the subject:
The traditional search method of using the precoordinated UDC numbers as queries is enriched by additional search possibilities represented by two types of descriptors in our experimental database namely: monolingual Romanian descriptors assigned by the indexers at the moment of subject indexing and multilingual descriptors in English, French and Romanian assigned automatically. Likewise, the UDC numbers in the classified catalogue decomposed into their component parts and particularly their textual meaning can be successfully used in searching.

3.3 Consequences of the automatic indexing on information retrieval

The search we present as an example used queries in Romanian and in French and it was first conducted by using automatically assigned multilingual descriptors in a) and b) and then by manually assigned Romanian ones existing in the bibliographic database in c). A comparison between these search results will show the higher recall rate of automatically assigned multilingual descriptors used as queries as against the manually assigned ones [see below the responses to the search statement ‘Cardiovascular diseases’: 101 hits for 1 query statement vs. 80 hits for 6 query statements].

The results are quite promising in as much as information retrieval is concerned if we compare the search results for a given topic using each of the two search possibilities in turns, i.e. the manually assigned monolingual descriptors and the automatically assigned multilingual ones.
Consider another example. It is often difficult to pinpoint exactly the right term for a particular concept. Let us consider the following UDC numbers in the MRF and see what the consequences are when they are used in indexing and mapped (manually) onto descriptors.


159.923.3 Composition of the personality. Character traits. Psychogram

159.925 Study of expression. Physical expression of character

For the first UDC number there are 3 different descriptors used in the bibliographic records in the database: Tipologie (Psihologie), Psihologie individuala and Caracter (Psihologie).

The second UDC number is represented in the bibliographic records by descriptors such as: Personalitate (Psihologie), Caracter (Psihologie), Caracterologie, Tipologic (Psihologie).

The third UDC number has as textual correspondents the following terms: Psihologie individuala, Fizionomie, Caracter (Psihologie), Morfopsihologie - here it was the title of the document that inspired the use of this descriptor: ABC de la morphopsychologie : [connaître sa personnalité par les traits du visage] / Carleen Binet.

Needless to say that the diversity of manually assigned descriptors can only produce confusion because of total inconsistency in indexing but the final result is loss of information having as source information scattering throughout the catalogue. The necessity of control on terms is imperious here and one way to
provide it is to make cross references between descriptors based on the principle of likeness. A much 'safer' way to ensure consistency in indexing is to map unique descriptors onto UDC notations. This way the control on terms is (automatically) imposed and this is just what our approach is about.

One of the requirements of the above-described approach is that there has to be a close relationship between the specificity of the classified catalogue of the given database and the specificity of the UDC-based thesaurus for better evaluating the system. If this request is ignored then we have a problem of compatibility between the classified catalogue and the selection of UDC numbers the thesaurus is based on. The direct consequence of this situation is information loss as in the following example. The main topic represented in the existing UDC notation 616.281-089.843 is too detailed compared with the general level of specificity of the UDC-based thesaurus. However, in a system like this there are ways to avoid such shortcomings given the alternative IR methods which can be used. One of those is the textual counterpart of the UDC notation itself (fields 702-709 in Figure 2).

Looking back at the table presenting the structure of the experimental database we can easily notice the difference between the number of records in the bibliographic database (97922) and the number of bibliographic records indexed with automatically assigned multilingual descriptors (84110). Presumably all bibliographic records are indexed with UDC numbers. The automatic indexing resulted in a lower number of records indexed with multilingual descriptors. The reason has to do with the difference in specificity between the two indexing languages i.e. the classification notations found in the database and the selection of classification numbers used as a base for the multilingual thesaurus.

4. Advantages of a multilingual UDC-based thesaurus

These examples are meant to argue the following:

- Being given a bibliographic database with subjects indexed by UDC numbers and descriptors assigned manually according to the meanings of those UDC numbers, the automatically assigned descriptors derived from a UDC-based thesaurus turn out to give better results in information retrieval. Therefore the improved retrievability is the main advantage of our approach.

- The multilingual character of the UDC-based thesaurus implies automatically that it addresses a broader range of users according to the number of languages involved.

- The manually assigned descriptors are not so reliable as the automatically assigned ones given the inconsistencies and lack of control likely to occur when more than one descriptor is linked with the same UDC number; the UDC-based descriptors will always be the same for a particular UDC number. It is the number of lead-in terms in each of the contributing languages that provides lexical richness to the indexing language anticipating the users' formulations therefore enhancing predictability.
Figure 2. Bibliographic record showing the complementarity of the alternative indexing languages

- The user-friendliness of this indexing tool is beyond any doubt. The accurate use of classification notations combined with the necessary respect of the UDC grammar rules in the classifying process are the basic requirements for an optimal result in the automatic conversion of the UDC notations and the assignment of the descriptors from the multilingual UDC-based thesaurus. Such a way the highly preferable natural language words replace the inconvenient numeric codes in information retrieval.

- The automatic assignment of UDC-based descriptors in the bibliographic records has an extra advantage derived from the user-friendliness of this indexing method. If numbers do not say much when it comes to appreciating the correctness of a given UDC notation, a descriptor that is automatically converting that notation will immediately struck the eye if not correctly assigned. An apparently minor typing mistake can tremendously distort the subject of a document affecting thus the recall rate. Example: A switch
between two digits in the UDC notation: 239.18 Dogmatic theology instead of 329.18 Fascist attitude for a document having the title "The last days of Hitler" may at least generate confusion in the searcher's mind. Such an error is much easier to be identified in natural language words than in classification codes. Under such circumstances the keeping up of the database will demand less efforts particularly if global change procedures are available for that purpose.

Our approach illustrates how traditional resources can be integrated into newly-designed models of knowledge organizers by mapping terms in several languages onto the same concept expressed in a language-independent structure. The UDC in our case plays a creative role in information retrieval participating in new strategies for improved access to information in multilingual environment.

Reference
Annelise Mark Pejtersen and Hanne Albrechtsen
Centre for Human-Computer Interaction
Risø National Laboratory, Denmark

Models for Collaborative Integration of Knowledge

Abstract: Collaborative integration of knowledge in distributed and cross-disciplinary work domains poses a number of challenges to classification, comprising: 1) how to analyze the actors' current practice of integration of knowledge and 2) how to model consistent semantic support of diverse interpretive perspectives among the actors. This paper introduces a cognitive systems engineering approach to modelling collaborative integration of knowledge in work domains. A generic means-ends model provides a theoretical foundation for mapping the territory of collaborative work. A decision task model captures the actors' distributed decision-making in integration of knowledge. The problem of collaborative integration of knowledge in a distributed web-based film collaboratory is explored through an empirical case of collaborative film indexing. The empirical study identified a lack of tools for consistent support of integration of knowledge. The means-ends model and the decision task model guided the design of a conceptual structure of the common workspace of film indexing. The paper concludes with a proposal for further work on models for integration of knowledge through ecological classification schemes.

1. Introduction

Current work practice and knowledge production to an increasing degree involves actors from different disciplines, cultures and organisations. Additionally, current work practice not only relies on authoritative orderings of knowledge, but also relies on the dynamism of the actors' ongoing collaborative integration of knowledge, i.e. their shared interpretations of knowledge, exchange of perspectives and joint knowledge production. Consequently, in order to support the actors' ongoing collaborative integration of knowledge, the design of support tools, like classification schemes, must address not only the order of knowledge, but also the situational contexts where collaborative integration of knowledge occurs. This paper introduces an ecological approach to integration of knowledge across boundaries in distributed collaboratory work environments, which is founded on (a) work domain analysis (b) the development of models for collaborative integration of knowledge. The work domain analysis is based on means-ends analysis of the territory of work and the actors' information needs during decision making. The result is conceptual structures of collaborative work that can be used to create collaborative classification schemes. Previous work on design of ecological classification schemes proposed that such schemes should be based on a fine-grained empirical analysis of actors' collaborative decision tasks in order to identify the knowledge produced and needed by the actors (Pejtersen & Albrechtsen, 2000).
2. Related work
Integration of knowledge in work domains through classification schemes has gained increasing interest among a variety of disciplinary fields and research communities. These comprise social studies of science and social anthropology, computer science, information science and the interdisciplinary field of computer-supported collaborative work (CSCW). Contributions by social studies of science consider the intentional (social, cultural, organizational) factors that shape the integration of organizational or disciplinary knowledge, as inscribed in for instance classification schemes (Star and Griesemer, 1989; Law, 1994;). Within computer science, research on ontology engineering addresses organizational requirements for integration of knowledge in the design and use of company ontologies (Schmitz-Esser, 2000; Gruber, 1995). Within information science, a number of recent contributions mobilize new design ideas for integration of knowledge across boundaries that are grounded in the evolving practice of knowledge domains (Peffersen and Albrechtsen, 2000; Bowker and Star, 1999; Hjørland and Albrechtsen, 1999; Huber and Gillaspy, 1998). Integration of knowledge through classification schemes has also recently entered the arena of research discussion and knowledge production within the field of computer-supported cooperative work (CSCW), resulting in ideas for development of computer-based tools for actors' distributed design and maintenance of classification schemes for coordination of work (eg. Schmidt and Wagner, 2001; Simone and Narini, 2001). The diversity of explicit as well as implicit methodological foundations for classification schemes to integrate distributed domain knowledge is addressed by Albrechtsen (2002), proposing a reconceptualization of general design principles for classification in order to develop a conceptual framework for the design of ecological classification schemes.

3. Generic models for work domain analysis
One of the aims of the ecological approach is to develop models that can support the development of ecological collaborative classification schemes for integration of knowledge in distributed collaboratory work environments. Two generic models guide the formulation of collaborative classification models: (i) the generic means-ends model (figure 1) (i) the generic decision task model (figure 2).

3.1 Generic means-ends model
The first model (figure 1) guides analysis of the territory of work in terms of the means and ends of work domains. The aim of this work domain analysis is to identify the stable means and ends that underlie the responses of the collaborating actors. The stable means and ends of work shape the behavior of actors, and they guide their activities. This model considers any unit of work as a functionally coupled entity, which adapts to the opportunities and requirements addressed by the actors. The aim is not to represent the actual coupling and functioning of the territory of work in particular situations. The purpose is instead to produce a generalized representation of the territory of work. This representation concerns categories of goals, functions, activities, and physical resources - all of which constitute the territory in which the actors work. This representation includes both general properties of the territory of work and the specific properties, which can be used to model the possibilities for action as well as the limitations that can constrain these actions.
3.2 Generic decision task model

However, the means ends model is a stationary, situation-independent representation of the territory of work. The actors will need to deal with and activate parts of this territory during particular decision tasks that arise in various work situations. In any particular decision task, a relevant activity space constitutes the basis for the activities of the involved actors. This activity space will include only a limited set of means ends relationships. To be able to relate the relevant set of means ends relationships to the actors' recurrent decision tasks, it is necessary to identify the decisions that are active in collaborative task situations. Decision tasks must be decomposed into sub-tasks and analyzed. The focus of this analysis is on (i) the information about means and ends that is communicated among actors, and (ii) the resulting knowledge that the actors arrive at about the state of affairs, as well as their plans for changing the current state of affairs. Basically different decision processes are required to be able to relate – and possibly integrate - the actors' levels of knowledge about the current state of affairs. Each of these decision processes leads to different kinds of knowledge sharing and integration. There are three generic types of decision processes:

Situation analysis

The situation analysis is initiated by the activation of the actors' attention, and it involves observations and questions, problem identification, exchange of perspectives, identification of options, which are all analytical processes.

Evaluation of options

The implications of the actual state of affairs with reference to the current goals and constraints must be evaluated. This analytical process involves prediction, value judgement, possible reassessments and the actors' priority making when considering choice among options and possible actions.

Planning actions

Based on the analysis of the state of affairs and choice among possible solutions to the identified problems, the proper sequence of actions must be selected. This evolves around the process of planning and scheduling and involves decisions of actors and work situations that will be engaged in the execution of a decided action.

The decision task model in figure 2 represents the set of generic sub-tasks involved in decision-making. The model is based on analyses of verbal protocols of actual work in different work domains. These protocols identified the different categories of statements made by the actors about their knowledge, their questions regarding their tasks and other related tasks and actors, as well as their past and intended acts (Rasmussen, Pejtersen & Goodstein, 1994).

The decision processes involved in actual work are not structured in rational sequences according to situation analysis, evaluation of options and planning of actions. Actors iterate among the different processes according to their levels and perspectives of knowledge. The model (figure 2) represents the various levels and perspectives of knowledge and the kind of information that the actors require about the territory of work in order to go from one level of knowledge to another. Identification of the knowledge levels and different domain perspectives of collaborating actors in joint decision situations is important for modelling decision tasks in collaborative task situations.
4. Integration of knowledge through a web-based collaboratory

The problem of collaborative integration of knowledge is presently being addressed by the European COLLATE project (Pejtersen et al., 2001) on the creation of web-based collaboratories for integration of knowledge in film research among European film archives. Currently, numerous valuable historic and cultural films and their sources are scattered in various national archives in Europe. To fully exploit the cultural film heritage internationally, a high degree of cross-disciplinary collaboration and integration of knowledge among actors working with the film media is required. The Collate project is at present investigating how collaborative research, cultural mediation and preservation of films can be supported through a distributed multimedia web collaboratory. One of the primary aims of the Collate project is to develop compatible methods and tools for ongoing collaborative integration of knowledge on the web. These methods and tools will support interactive discussions and shared interpretations of films, interactive coordination of preservation activities as well as collaborative classification, indexing and annotation of films. The current collaborative integration of knowledge among staff within and among the archives involved in the Collate project was explored in field studies and empirical work analysis, guided by the framework for cognitive systems engineering (Rasmussen, Pejtersen & Goodstein, 1994). The Collate project is presently designing prototypes for collaborative film indexing and annotation. The following empirical study carried out within the Collate project addressed the case of integration of knowledge through a means-ends analysis and decision task analysis of collaborative film indexing in a national film archive.

4.1 Supporting integration of knowledge in collaborative film indexing

The national film archive accords national film-production high cultural value (Pejtersen et al., 2001). This is reflected in its goals of collecting, archiving, preserving and restoring all national film-related materials. Special attention and effort are directed to uniform description (indexing and cataloging) of all materials of archive's collections and to preserving and restoring disappearing knowledge. This focus is supported through the archive's collaboration with national film experts, as a means of resurrecting historic film material. The archive's collaborative film research is documented in a national filmography. The production of the filmography involves collaboration on film indexing and description among staff within the archive as well as with film experts outside the archive. The actors represent a multitude of cross-disciplinary perspectives on film research, for instance, film studies, drama studies, social history, restoration, education, women's studies. The empirical study of the archive reported below focuses on integration of knowledge through collaborative film indexing in two different work situations (Albrechtsen, Pejtersen & Cleal, 2002).

Collaborative work situations and decision making.

Two different collaborative work situations take place at the archive: (1) screening meeting for collaborative analysis of films (2) collaborative film description. The goal of these two work situations is to create and integrate knowledge about films:

Screening meeting

The screening meeting involves film experts outside the archive together with archive staff, engaged in filmography production. The main purpose of the meeting is to elicit knowledge from all actors on film subjects, and to bring out different domain perspectives on these subjects and possible changes in interpretations of a film across time. Another purpose is assessment of the relevance of a film. The
common workspace is comprised of the meeting room, a film and an agenda for the meeting. The primary work object for the group discussion is the film itself.

Film description

Following the screening meeting, the staff (filmographers) engages in collaborative film description. The individual filmographers informally consult one another and the experts that participated in a screening meeting. The main purpose of this latter collaborative work arrangement is to enter information about a film in a database for subsequent publication in the printed national filmography. The common workspace consists of the database structure, the filmography, the archive’s collections and standard rules for film description (cataloging). The primary work object is still the film itself. However, the filmography staff’s attention focus shifts to the creation of filmography entries.

Each of the two work situations, i.e. (1) screening meeting and (2) film description was analyzed by use of the decision task model (figure 2). Together, the two work situation were found to comprise the following overall task decisions for collaborative film indexing:

<table>
<thead>
<tr>
<th>Screening meeting:</th>
<th>Film description:</th>
</tr>
</thead>
<tbody>
<tr>
<td>screening and relevance assessment (evaluation decision task)</td>
<td>content description (renewed evaluation decision task)</td>
</tr>
<tr>
<td>content analysis of film (situation analysis decision task)</td>
<td>planning filmographic entries and entering data</td>
</tr>
<tr>
<td>(planning decision task)</td>
<td>(planning decision task)</td>
</tr>
</tbody>
</table>

The transition from the content analysis that takes place during the screening meeting to the content description in the film description task involved substantial revision of film interpretations given at screening meetings. This is among other issues due to the interpretive regulations imposed by the current work tools (database, indexing terms) and codes of practice (cataloging rules) applied by the archive staff. These factors were found to constrain the renewed evaluation decision task. The purpose of the applied tools and codes of practice was to support the actors in coping with the conceptual complexity of translating film interpretations from screening meetings to filmography entries in a uniform and standardized way. As a result, the multiple perspectives of film interpretation elicited at screening meetings were only partially accounted for in the filmography entries. Thus, high level interpretations made by the participants in the screening meeting about the intentions and goals by film originators were not entered into the filmography and consequently not available for subject searching in the filmography database.

Support of collaborative film indexing through a common conceptual structure.

Despite the shift of work from screening meeting to film description and the change of actors’ attention focus, the common work problem is the same: to assess, analyze and describe a film for a national filmography. In order to provide continual and consistent semantic support of integration of knowledge during collaborative indexing, a conceptual means-ends structure of the collaborative workspace was created (figure 3). The design of the structure was guided by the means-ends model
(figure 1) and the AMP scheme for fiction mediation (Pejtersen, 1994). The conceptual structure of the collaborative workspace for film indexing (figure 3) covers all levels of film interpretations and characterizations from overall intentions and goals to archival attributes of distribution, provenance and mediation. This structure can make visible the state of work throughout collaborative indexing activities, from screening meeting to production of filmography entries. The conceptual structure is pertinent for screening meetings and for taking notes on decisions made by the actors at the meeting. Furthermore, it is relevant for continual backtracking and reassessment of filmographic entries. In addition, the conceptual structure can support concurrent collaborative integration of film knowledge where actors that have participated in a screening meeting may interactively annotate filmography entries. Finally, the conceptual structure is applicable for designing an ecological classification scheme where the dimensions of the structure are relevant as fundamental divisions for semantic structuring of the archive's alphabetical list of indexing terms.

Conclusion and future work

The model of the common conceptual workspace of film indexing was part of Risø's contribution to the Collate project on the creation of a web-based multimedia collaboratory for integration of knowledge in European film research. The problem of integration of knowledge in the cross-disciplinary domain of film research was addressed through empirical work analysis, means-ends analysis and decision task analysis. Based on these analyses, a common conceptual structure of the collaborative workspace was created. Future work on collaborative ecological classification schemes will model the coupling of decision activities that are shared among distributed collaborating actors within a means-ends structure. This coupling of decision activities can occur within one institution or across institutions. An important aim of this modelling is to design collaborative ecological classification schemes that articulate the ongoing production and integration of knowledge, as elicited during collaborative decision-making, through an ordered exposition of integrated domain knowledge.

Acknowledgements

We gratefully acknowledge the support of the Centre for Human-Machine Interaction (CHMI), Risø National Laboratory, Denmark, and the Danish National Research Foundation who funded the research presented in this paper. We also acknowledge the support of the European Commission's IST programme that funded the Collate project. In addition, we thank the partners of the Collate Consortium, FhG-IPSI, Darmstadt, Germany, Deutsches Filminstitut, Frankfurt, Germany, Filmarchiv Austria, Vienna, Austria, Národní Filmovy Archiv, Prague, Czech Republic, University of Bari, Laboratory LACAM, Bari, Italy and Sword ICT, Bari, Italy.

References


<table>
<thead>
<tr>
<th>MEANS-ENDS RELATIONS</th>
<th>PROPERTIES REPRESENTED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goals and Constraints</td>
<td>Properties necessary and sufficient to establish relations between the performance of the system and the reasons for its design, i.e., the purposes and constraints of its coupling to the environment. <em>Categories are in terms referring to properties of environment.</em></td>
</tr>
<tr>
<td>Priority measures</td>
<td>Properties necessary and sufficient to establish priorities according to the intention behind design and operation: Topology of flow and accumulation of mass, energy, information, people, monetary value. <em>Categories in abstract terms, referring neither to system nor environment.</em></td>
</tr>
<tr>
<td>General Functions</td>
<td>Properties necessary and sufficient to identify the ‘functions’ which are to be coordinated irrespective of their underlying physical processes. <em>Categories according to recurrent, familiar input-output relationships.</em></td>
</tr>
<tr>
<td>Physical work Processes and Activities</td>
<td>Properties necessary and sufficient for control of physical work activities and use of equipment: To adjust operation to match specifications or limits; to predict response to control actions; to maintain and repair equipment. <em>Categories according to underlying physical processes and equipment.</em></td>
</tr>
<tr>
<td>Physical resources and Configuration</td>
<td>Properties necessary and sufficient for classification, identification and recognition of particular material objects and their configuration; for navigation in the system. <em>Categories in terms of objects, their appearance and location.</em></td>
</tr>
</tbody>
</table>

Figure 1: The means-ends abstraction hierarchy
Figure 2: Generic decision task model. The figure should be read from the left column situation analysis from the bottom beginning with activation of attention, then followed by the evaluation of goals and options, and then down to the right column of planning actions.
1. Originators' goals and values

**Why?**  
(1) **Originators' intentions and goals:** Communication of information; education; promotion and stimulation of ideas and emotional, socio-cultural and aesthetic experience  
**Constraints:** censorship, production means

**Why?**  
(2) **Originators' affiliation and attitude:** Professional paradigms (narrative/aesthetic style, form, etc.) and their socio-historical context, political and cultural movements, or other value criteria

2. General and specific content in film

**When?**  
(3) **General frame/time content:** Time, year, historical period

**Where?**  
(4) **General frame/place content:** Place and setting, geographical, historical, socio-cultural contexts of the topic

**What?**  
(5) **Subject matter content:** Specific topic and plots; psychological and social phenomena; ending of film

**Who?**  
(6) **Living beings, institutions and artefacts:** Main characters, persons, animals, plants, institutions that are involved in the topic, cast

3. Communication and presentation of film format

**How?**  
(7) **Film types and formats**

**How?**  
(8) **Accessibility level,** eg. censorship documentation/cards, contemporary documentation, sources, bibliography (reviews etc.), indexes

**How?**  
(9) **Physical characteristics:** eg. sound systems

4. Filmography attributes

**Who?**  
(10) **Filmographic data and identification numbers**

**Where?**  
Locations

**When?**  
Year of production, version

**What?**  
Type of material

**How?**  
Size and format

5. Archive attributes

**Why?**  
(11) **Archive goals and policies**

(12) **Local archival conditions:** eg. shelving, storage

**When?**  
Preservation, restoration, sales and lending of films and videos

6. Distribution, provenance, mediation

**Why?**  
(13) **Distributors', donators' etc. goals and policies:**

**Where?**  
(14) **Production and film studios**

**When?**  
(15) **Premieres (dates), awards**

Figure 3: Conceptual structure of the common workspace of collaborative film indexing
Evaluation of the Application of Genetic Algorithms to Relevance Feedback

Abstract: We evaluated the different genetic algorithms applied to relevance feedback that are to be found in the literature and which follow the vector space model (the most commonly used model in this type of application). They were compared with a traditional relevance feedback algorithm - the Ide dec-hi method - since this had given the best results in the study of Salton & Buckley (1990) on this subject. The experiment was performed on the Cranfield collection, and the different algorithms were evaluated using the residual collection method (one of the most suitable methods for evaluating relevance feedback techniques).

The results varied greatly depending on the fitness function that was used, from no improvement in some of the genetic algorithms, to a more than 127% improvement with one algorithm, surpassing even the traditional Ide dec-hi method. One can therefore conclude that genetic algorithms show great promise as an aid to implementing a truly effective information retrieval system.

1. Introduction

We all know that even the best information retrieval system (IRS) presents limitations in recall: the user hardly ever retrieves all the documents that are relevant to his or her information requirements.

One possible approach to this problem is for the IRS to offer the user the possibility of modifying the original query by means of relevance feedback, the most popular of the query modification strategies. Introduced in the 1960s, this is a controlled and automated process of query definition, which is both simple to use and extraordinarily effective. The main idea is to choose important terms linked to certain previously retrieved documents which the user has identified as relevant, and to increase the importance of these terms in the new formulation of the query. Similarly, one can decrease the importance of the terms included in irrelevant documents that were retrieved previously (Salton & McGill, 1983).

There have been many studies on this subject, since the advantages of relevance feedback together with its relative simplicity to implement, make it a highly desirable technique to include in any IRS. A landmark work in this line was that of Salton & Buckley (1990). This compared the most commonly used traditional methods of relevance feedback applied to six test collections, including the Cranfield collection. In most cases, there were major improvements in retrieval
with respect to the original query. The most efficacious was the Ide dec-hi method (Ide, 1971).

Also, recent years have seen the appearance of applications of genetic algorithms (GAs) to information retrieval. Most of these applications use the vector space model, and fall into three main groups according to their field of application: document indexing, clustering, and relevance feedback (Cordon et al., 1999), with the last of the three being the most numerous group.

GAs are artificial intelligence techniques which use adaptive procedures, aimed at solving search problems that satisfy certain prerequisites. Their inspiration lies in the mechanisms of biological evolution based on natural selection, and in the genetic codes of species (Holland, 1992; Michalewicz, 1995).

An application of GAs to relevance feedback may start from a set of possible solutions (the queries) which the algorithm causes to evolve with the aim of optimizing the solution. To guide the evolution of the possible solutions, one has to design fitness functions that allow one to evaluate the goodness of each solution on the basis of feedback from the user.

There has not as yet, however, been any comparison of the results obtained in isolation in this type of experiment with the traditional relevance feedback algorithms. This will be the main objective of the present study - to evaluate the different GAs applied to relevance feedback that are to be found in the literature, and which follow the vector space model, using the Cranfield collection and comparing the results with the classical Ide dec-hi method.

2. Antecedents

For obvious reasons, we can not here give an exhaustive description of the different applications of GAs to relevance feedback. We shall only note that most of them use a very similar and basic genetic algorithm, with the main difference between one and another method lying in the fitness function that is used. Hence we shall briefly comment on these different functions, which we shall later implement in an optimal feedback genetic algorithm.

Robertson & Willett (1996) calculate an individual's fitness by first determining the inner product of each query of the collection with each document of the database. They then retrieve a fixed number of documents, and finally calculate the recall of the retrieval (the functions fitness 1 and fitness 2, with the latter being a variant of fitness 1 using a cosine similarity measure instead of the inner product).

The type of fitness function used by Chen's group (Chen, 1995; Chen & Iyer, 1998; Chen et al., 1998) consists in determining for each chromosome (possible query) the Jaccard index with the rest of the population's chromosomes. They then obtain the mean value of these similarities (the functions fitness 3, fitness 4, and fitness 5, with the last two being practically the same as the first but using the inner product and the cosine, respectively).

To calculate each individual's fitness, Yang & Korfhage (1994) compare each query with the collection's documents using the Euclidean distance, retrieving those documents whose distance is less than a certain threshold. The fitness is obtained with the following expression:

\[ F_i = Rr - Rn - Nr \]

where \( Rr \) is the number of documents retrieved that belong to the standard set of relevant documents provided by the collection, \( Rn \) is the number of retrieved
documents that do not belong to the standard set of relevant documents, and \( N_r \) is the number of relevant documents that were not retrieved (the functions \( \text{fitness 6} \) and \( \text{fitness 7} \), where the latter is similar to \( \text{fitness 6} \), but, instead of calculating a threshold for the retrieval, the first ten documents are considered).

A most innovative work from the perspective of the fitness function is that of Horng & Yeh (2000). As well as the number of relevant and of irrelevant documents retrieved, the fitness function takes the order in which they appear into account, since it is not the same whether the relevant documents are given at the top or at the bottom of the list of retrieved documents. Their function is constructed as follows: One calculates the similarity of the query vector with all the documents using the inner product, ranks the documents in order of decreasing similarity, and finally calculates the fitness of the chromosome with the following expression:

\[
F = \frac{1}{|D|} \sum_{i=1}^{[D]} r(d_i) \sum_{j=i}^{[D]} \frac{1}{j}
\]

where \( D \) is the total number of retrieved documents, and \( r(d) \) is a function giving the relevance of document \( d \), being unity if the document is relevant, and zero otherwise (the function \( \text{fitness 8} \)).

As well as an experiment with a fuzzy representation, Martín-Bautista (2000) performs another experiment with a real representation, using for the fitness function the following combination of the classical measures of precision and recall:

\[
F = \nu \cdot \text{recall} + (1 - \nu) \cdot \text{precision},
\]

setting \( \nu \) to 0.4 (the function \( \text{fitness 9} \)).

In López-Pujalte et al. (2000, 2002a), we worked with two fitness functions which took the order of appearance of the documents into account, as in Horng & Yeh. For the first of the functions (\( \text{fitness 10} \)), we made a cosine comparison of the indicated chromosome and the feedback documents. After ranking the documents by relevance, we made a cumulative calculation for all the retrieved documents in the following form:

\[
T = \sum_{pos=1}^{num} r(d_i) \frac{1}{A} \left( \frac{A - 1}{A} \right)^{pos-1}
\]

where \( pos \) is the position of the document under consideration, \( num \) the total number of documents retrieved, and \( r(d) \) a function returning the relevance of document \( d \), being +1 if the document is relevant, and -1 otherwise. The contribution will therefore be positive for a relevant document and negative for an irrelevant document. Finally, the accumulated total result is multiplied by the recall of the retrieval to give the chromosome's fitness. The optimal value of the parameter \( A \) was fixed by experiment to 2.0.

The second of our functions (\( \text{fitness 11} \)) is calculated as follows: First, the cosine method is used to calculate the similarity between the indicated chromosome and the feedback documents. They are then ranked into decreasing degree of similarity, and the mean precision is calculated in nine recall intervals (0.1, 0.2, ..., 0.9), including an interpolation procedure to remove ambiguities.
We can not bring this section to a close without mentioning the Ide dec-hi method (Ide, 1971), with which we shall compare the GAs. This method consists in simply appending to the terms of the original query the terms of all the relevant documents in the set of feedback documents (15 in our case), and removing from the terms those corresponding to the first irrelevant document of the feedback set obtained in the retrieval.

3. Experiment and evaluation

To perform our trials, we needed to generate a testbed database, which we created from the Cranfield collection. We took all the 1398 documents of the collection, and a subset of 33 of the queries (these were those which are appropriate to testing the relevance feedback technique: not all the queries in the collection are suitable).

Then, to determine the terms that we would use to describe the documents of the collection, we performed a procedure of extracting words from each document of the database, eliminating the stopwords, and stemming the rest. This left 4307 terms, so that we worked with 1398 4307-dimensional vectors.

To assign the weights, we used the scheme put forward by Salton & Buckley (1990) so as to equiparate our experiment as closely as possible to theirs. We finally normalized the vectors.

We carried out a similar procedure with the collection's associated queries, thereby obtaining the normalized query vectors which will be the object of relevance feedback optimization.

The next step after this vectorization process was to carry out the trials. We finally decided on 15 as the number of documents with which to implement the feedback. I.e., for each query, the first 15 documents retrieved were examined (using the cosine method) for their relevance, and this information was fed into the algorithm as feedback.

With respect to the genetic methods, the genetic algorithm that was finally used for the relevance feedback was the one that performed best amongst all the alternatives that were tried (López-Pujalte, 2000; López-Pujalte et al., 2002b). It has the following characteristics:

- The chromosomes use real coding, and have as many genes as there are non-zero weights in the query and in the 15 documents that the system will use as feedback (Chen, 1995).

- It is a hybrid genetic algorithm, which receives as the initial population the chromosomes corresponding to the relevant documents, the irrelevant documents, the latter with their terms negated (in all cases referring to the documents of the feedback set), and lastly to the supposedly optimized query.

- The genetic algorithm uses simple random sampling with elitism as selection mechanism, and simple one-point crossover and random mutation as genetic operators. Before applying the genetic operators, the population's chromosomes are normalized, since this was found to significantly improve the genetic algorithm's behaviour.

- On exiting, the genetic algorithm returns as solution both the best chromosome found during the process and the centroid of the final population chromosomes whose fitness is equal to the maximum value found in that generation. Both solutions were implemented and evaluated.
- The control parameters were set experimentally at 0.8 for the crossover probability, 0.2 for mutation, and 20 for the maximum number of generations.

We evaluated the results of the retrieval via the classical measures of recall and precision, with the latter calculated by interpolation at fixed recall intervals as described by Salton & McGill (1983). We also used the residual collection method, in which all the documents previously seen by the user (whether relevant or not) are removed from the collection, and both the initial and the modified query are evaluated on this residual collection (Salton & Buckley, 1990).

4. Results

The final results are listed in Table 1, which gives the mean precision for each algorithm implemented, as well as the percentage improvement with respect to the unoptimized initial query.

One sees that the results vary greatly depending on which fitness function was used - from no improvement with the fitness functions introduced by Chen's group to an improvement of 127% with fitness function 11.

One sees therefore that the relevance feedback technique which yields the best results is a genetic technique, specifically that which uses fitness function 11 that we ourselves had designed. This performs better than both the rest of the genetic techniques and the classical Ide dec-hi method.

With respect to the calculation of the solution, in most cases the genetic algorithms yielded the best results with the first method of finding the solution ("the best"), except for the GAs which use fitness functions 8 and 11, where the second method ("centroid") gave the best results.

<table>
<thead>
<tr>
<th>METHOD</th>
<th>THE BEST</th>
<th>CENTROID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MEAN</td>
<td>IMPROVE</td>
</tr>
<tr>
<td>Without feedback</td>
<td>0.098</td>
<td></td>
</tr>
<tr>
<td>Ide (dec-hi)</td>
<td>0.218</td>
<td>120.8 %</td>
</tr>
<tr>
<td>GENETIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GA with Fitness 1</td>
<td>0.218</td>
<td>120.8 %</td>
</tr>
<tr>
<td>GA with Fitness 2</td>
<td>0.218</td>
<td>120.8 %</td>
</tr>
<tr>
<td>GA with Fitness 3</td>
<td>0.043</td>
<td>—</td>
</tr>
<tr>
<td>GA with Fitness 4</td>
<td>0.060</td>
<td>—</td>
</tr>
<tr>
<td>GA with Fitness 5</td>
<td>0.053</td>
<td>—</td>
</tr>
<tr>
<td>GA with Fitness 6</td>
<td>0.148</td>
<td>50.7 %</td>
</tr>
<tr>
<td>GA with Fitness 7</td>
<td>0.218</td>
<td>120.8 %</td>
</tr>
<tr>
<td>GA with Fitness 8</td>
<td>0.218</td>
<td>120.8 %</td>
</tr>
<tr>
<td>GA with Fitness 9</td>
<td>0.218</td>
<td>120.8 %</td>
</tr>
<tr>
<td>GA with Fitness</td>
<td>0.218</td>
<td>120.8 %</td>
</tr>
<tr>
<td>GA with Fitness</td>
<td>0.218</td>
<td>120.8 %</td>
</tr>
</tbody>
</table>

Table 1: Results of the different relevance feedback methods for the Cranfield collection.

With respect to runtimes, the Ide dec-hi method is decidedly faster. It uses only 0.12 seconds per query. The GAs, however, still have quite acceptable
runtimes of between 3 and 5 seconds (the best GA, for example, uses 3.39 seconds per query). The data processing equipment used in the experiments consisted of a Pentium III server running at 700 MHz with 256 MB of RAM, and a 26 GB hard disk.

5. Conclusions

One may deduce from the results that these algorithms allow one to considerably improve the original query by implementing the relevance feedback technique (in a single iteration), surpassing even the improvement obtained with the classical Ide dec-hi method. Also, the choice of fitness function to use is of prime importance, since some functions can lead to success and others to utter failure in the exploration.

Amongst the functions that we tested, the best results were given by those which take into account not only which documents are retrieved, but also the order in which they are retrieved. I.e., it is advisable to use fitness functions which value not only that the possible solution retrieves many relevant and few irrelevant documents, but also that the relevant documents are at the top of the list and not at the bottom.

References


An Inductive Query by Example Technique for Extended Boolean Queries Based on Simulated-Annealing Programming

Abstract: One of the key problems that non-expert users have to deal with when using an Information Retrieval System is the need to deeply know its query language in order to express their information needs in the form of a valid query allowing them to retrieve relevant information. To solve this problem, Inductive Query by Example techniques can be considered to automatically derive queries from a set of relevant documents provided by a user. In this paper, a new hybrid evolutionary technique is proposed to automatically learn extended Boolean queries and is compared to Kraft et al.'s approach in several queries of the well known Cranfield collection.

1. Introduction

Information retrieval (IR) may be defined, in general, as the problem of the selection of documentary information from storage in response to search questions provided by an user (Salton & McGill, 1984). Information retrieval systems (IRSs) are a kind of information system that deal with data bases composed of information items —documents that may consist of textual, pictorial or vocal information— and process user queries trying to allow the user to access to relevant information in an appropriate time interval.

Most of the commercial IRSs are based on the Boolean model (van Rijsbergen, 1979), which presents some limitations. Due to this fact, some paradigms have been designed to extend this retrieval model and overcome its problems, such as the vector space (Salton & McGill, 1984) or the fuzzy information retrieval (FIR) models (Bordogna et al., 1995).

However, the increase in the power of the retrieval model also comes with a high complexity augment in the query language, which makes it difficult to represent the user information needs in the form of a valid query. This is especially significant in fuzzy IRSs: if it is difficult for a user to formulate a classical Boolean query due to the need to know how to properly connect the query terms together using the
Boolean operators, it is even more complex to both define the query structure and specify the query term weights to retrieve the desired documents.

Hence, the paradigm of Inductive Query by Example (IQBE) (Chen et al., 1998), where a query describing the information contents of a set of documents provided by a user is automatically derived, can be useful to solve this problem and assist the user in the query formulation process. Focusing on the FIR model, the most known existing approach is that of Kraft et al.'s (1997), which is based on genetic programming (GP) (Koza, 1992).

In this paper, a new IQBE technique for FIRSs based on a hybrid simulated annealing-genetic programming evolutionary algorithm will be introduced with the aim of improving the performance of Kraft et al.'s proposal in terms of retrieval accuracy. To do so, the paper is structured as follows. Section 2 is devoted to the preliminaries, briefly presenting the basis of FIRSs and of IQBE techniques. Then, Kraft et al.'s proposal is reviewed in Section 3. Section 4 presents the composition of the new algorithm proposed while the experiments developed to test it are showed in Section 5. Finally, several conclusions are pointed out in Section 6.

2. Preliminaries

2.1. Fuzzy Information Retrieval

FIRSs permits to deal with the uncertainty and imprecision existing in the retrieval process by extending classical Boolean IRSs in the three following aspects (Bordogna et al., 1995):

- Indexing terms do not absolutely describe (1) or do not describe at all (0) the document contents in the document representations, but they have a partial degree of aboutness in [0,1]. The indexing function \( F: D \times T \rightarrow [0,1] \) is a two-dimensional fuzzy set (a fuzzy relation), that is projected to obtain a fuzzy set associated to each document and term:
  \[
  d_i = \{ <t, \mu_{d_i}(t)> | t \in T \} ; \quad \mu_{d_i}(t) = F(d_i, t) \\
  t_j = \{ <d, \mu_{t_j}(d)> | d \in D \} ; \quad \mu_{t_j}(d) = F(d, t_j)
  \]

- Document retrieval also becomes a matter of degrees and the relevance of a document to a query is also measured in [0,1]. This allows the FIRS to rank the retrieved documents as regards their relevance to the query as in the vector space model.

- Finally, the query structure is extended by associating a numerical or linguistic weight to each query term. These weights can be interpreted in different ways, e.g., relative importance among the terms involved in the query can be appropriately expressed.

In this contribution we consider the latter interpretation for the query weights. For the operation mode of the corresponding FIRS matching mechanism, as well as for a description of the remaining two approaches, the interested reader can refer to Bordogna et al. (1995).

2.2. Inductive Query by Example

IQBE was proposed in (Chen et al., 1998) as "a process in which searchers provide sample documents (examples) and the algorithms induce (or learn) the key concepts in order to find other relevant documents". This way, IQBE is a process
for assisting the users in the query formulation process performed by machine learning methods (Mitchell, 1997). It works by taking a set of relevant (and optionally, non relevant documents) provided by a user—that can be obtained from a preliminary query or from a browsing process—and applying an off-line learning process to automatically generate a query describing the user's needs (as represented by the document set provided by him). The obtained query can then be run in the IRS to obtain more relevant documents. This way, there is no need that the user interacts with the process as in other query refinement techniques such as relevance feedback.

All of the machine learning methods considered in that paper (regression trees, genetic algorithms and simulated annealing) dealt with the vector space model. Besides, Smith and Smith (1997) proposed a Boolean query learning process based on GP. As regards the applications in FIRSs, we find the GP algorithm of Kraft et al. that will be reviewed in the next section and our niching GA-P method that extends the latter considering a more sophisticated evolutionary algorithm (Cordon et al., 2000). For descriptions of those of the previous techniques based on evolutionary algorithms refer to (Cordon et al., 1999).

3. The Kraft et al.'s IQBE Process for Extended Boolean Queries

The IQBE technique of Kraft et al. (1997) is a GP algorithm with the following composition:

- **Coding Scheme**: The fuzzy queries are encoded in expression trees, whose terminal nodes are query terms with their respective weights and whose inner nodes are the Boolean operators AND, OR or NOT.

- **Selection Scheme**: It is based on the classical generational scheme, where an intermediate population is created from the current one by means of Baker's stochastic universal sampling (Baker, 1987), together with the elitist selection.

- **Genetic Operators**: The usual GP crossover is considered (Koza, 1992), which is based on randomly selecting one edge in each parent and exchanging both subtrees from these edges between the both parents.

On the other hand, the following three possibilities are randomly selected —with the showed probability — for the GP mutation:

a) Random selection of an edge and random generation of a new subtree that substitutes the old one located in that edge (p=0.4).

b) Random change of a query term for another one, not present in the encoded query, but belonging to any relevant document (p=0.1).

c) Random change of the weight of a query term (p=0.5).

- **Fitness function**: The following function combining the classical precision and recall measures (for more information about them, see van Rijsbergen, 1979) is considered:

$$F = \alpha \cdot \frac{\sum d r_d \cdot f_d}{\sum d f_d} + \beta \cdot \frac{\sum d r_d \cdot f_d}{\sum d r_d}$$

with $r_d \in \{0,1\}$ being the relevance of document $d$ for the user and $f_d \in \{0,1\}$ being the retrieval of document $d$ in the processing of the current query. Moreover, as simple queries are always prefered by the user, a selection criterion has been
incorporated to the algorithm in order to consider more fitted those queries with a lesser complexity among a group of chromosomes with the same fitness value.

4. A New IQBE Process for Extended Boolean Queries Based on SA-P

Kraft et al.'s IQBE algorithm introduced in the previous section can perform well but it suffers from a key limitation of GP: it is very difficult to find proper values for the numerical weights as they are only altered by mutation during the evolution process. Hence, fuzzy queries with the optimal structure can be discarded by the selection procedure as the term weights involved in them are not well adjusted.

The latter problem can be solved by concurrently adapting both the query structure and the term weights (as done by the GA-P algorithm proposed in Cordón et al., 2000). In this paper, we do so by a hybrid evolutionary algorithm between Simulated Annealing (SA) and GP, the SA-P paradigm, proposed in Sánchez et al. (2001). The algorithm was based on encoding both a expressional part (the parse tree) and a value string (the coefficients involved in the expression) and adapting it within usual SA search scheme by a neighborhood operator based on the classical GP crossover and a string value mutation operator.

Hence, the extended Boolean query is encoded by storing the query structure —terms and logical operators— in the expressional part, and the term weights in the value string using a real coding scheme. The neighborhood operator (macromutation) generates the candidate fuzzy query by either changing the expresional part —the query structure— or the value string —the query weights— of the current individual \( I \). This decision is randomly made with respect to a value string mutation probability \( p \) (\( p=1 \) means "only weight mutation" while \( p=0 \) means "only query structure mutation"). The query structure is mutated by selecting an edge of its parse tree and substituting the subtree located at it by a randomly generated parse tree. The weight vector is mutated by applying intermediate recombination (Mühlenbein & Schlierkamp-Voosen, 1993) between the current values \( \text{weights}(I) \) and a randomly generated vector \( W \) with an amplitude parameter that depends on the current temperature \( T \) by a constant \( K_1 \) as follows: \( \text{weights}(I) = \text{weights}(I) \times T / K_1 + (1 - (T / K_1))W \).

The SA-P algorithm considered is showed as follows:

```
algorithm IQBE SA-P
needs: MaxEval /*maximum number of evaluations*/, MaxNeighs /*max. number of neighbors generated per temperature*/, MaxSuccess /*max. number of neighbors accepted per temperature*/, c /*cooling factor*/, T0 /*initial temperature*/, p /*value string mutation probability*/, K1 /*value string mutation parameter*/
produces: Ibest
I=Ibest=random individual; T=T0; Eval=1
while (Eval<=MaxEval) do
    num_neighs=num_success=0
    while (num_neighs<MaxNeighs) && num_success<MaxSuccess)
        Icand=macromutation(I,p,T,K_I);
        num_neighs=num_neighs+1
        delta=F(I)-F(Icand); Eval=Eval+1
return Ibest
```

\[ v = \text{random value with uniform distribution } U(0,1) \]

\[ \text{if } (\text{delta}<0) \text{ or } (v<\text{exp}(-\text{delta}/T)) \text{ then} \]

\[ I = I_{\text{cand}}; \quad \text{num\_success} = \text{num\_success} + 1 \]

\[ \text{if } (I>I_{\text{best}}) \text{ then } I_{\text{best}} = I \text{ end if} \]

\[ \text{end if} \]

\[ T = c^*T \]

\[ \text{end while} \]

As seen in the algorithm, the initial solution is a random fuzzy query. The initial temperature \( T_0 \) is computed by means of the following expression:

\[ T_0 = (\mu/\ln(\phi)) \cdot F(I), \]

with \( I \) being the initial solution and \( \phi \) being the probability of acceptance for a solution that can be \( \mu \) per 1 worse than \( F(I) \). Both parameters are defined in the interval \([0,1]\).

5. Experiments and Analysis of Results

We worked with the well known Cranfield documentary base to test the performance of our proposal. The 1400 documents were automatically indexed by first extracting the non-stop words, thus obtaining a total number of 3857 different indexing terms, and then using the normalized IDF scheme (Salto and McGill, 1984) to generate the term weights in the document representations. Among the 225 queries associated to this collection, we selected those presenting 20 or more relevant documents (queries 1, 2, 23, 73, 157, 220 and 225). The number of relevant documents associated to each of these seven queries are 29, 25, 33, 21, 40, 20 and 25, respectively.

Both our proposal and Kraft et al.'s algorithm were run on the previous relevant document sets. In order to make a fair comparison, both algorithms were run three times with different initializations during the same fixed number of fitness function evaluations (100000). For the sake of simplicity, only the experiments not considering the use of the NOT operator are reported (as done in (Kraft et al., 1997)).

The common parameter values considered are a maximum of 20 nodes for the expression parts, \((1.2, 0.8)\) for the fitness function weights \(\alpha\) and \(\beta\), and 0.1 for the FIRS retrieval threshold \(\phi\). Kraft et al.'s algorithm is run with a population of 1600 queries and 0.8 and 0.2 for the crossover and mutation probabilities respectively. Finally, for the SA-P algorithm, the initial temperature computation parameters \(\mu\) and \(\phi\) are set to 0.5, the maximum number of neighbors generated and accepted per temperature are respectively 500 and 50, the cooling parameter is set to 0.9, the value string mutation probability \(p\) takes value 0.5 and the parameter \(K_1\) considered for this mutation is set to 5.

The best result obtained by Kraft et al.'s and our method in each of the seven queries are respectively showed in Tables 2 and 3, where \#q refers to the query number, \textit{Run} stands for the corresponding algorithm run (1 to 3), \(T\) for the run time (both algorithms have been run in a 350 Mhz. Pentium II computer with 64 MB of memory, and the time is measured in minutes), \(S_z\) for the generated query size, \textit{Fit} for the fitness value, \(P\) and \(R\) for the precision and recall values, respectively, \#rt for the number of documents retrieved by the query, and \#rr for the number of relevant documents retrieved.
In view of these results, it is clear that our SA-P IQBE algorithm significantly outperforms Kraft et al.'s proposal in all the cases. The lower improvement corresponds to query 73 with a 37.5 percent (a recall value of 0.523810 for the SA-P against to 0.380952 resulting from Kraft et al.'s algorithm), and the highest one to query 157 with a 300 percent (a recall value of 0.225 for the SA-P against another of 0.075 got from Kraft et al.'s method).

<table>
<thead>
<tr>
<th>#q</th>
<th>Run</th>
<th>T</th>
<th>Sz</th>
<th>Fit</th>
<th>P</th>
<th>R</th>
<th>#rr/#rt</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,2</td>
<td>12:48</td>
<td>19</td>
<td>1.282759</td>
<td>1.000000</td>
<td>0.103448</td>
<td>3/3</td>
</tr>
<tr>
<td>2</td>
<td>1,2</td>
<td>12:44</td>
<td>17</td>
<td>1.328000</td>
<td>1.000000</td>
<td>0.160000</td>
<td>4/4</td>
</tr>
<tr>
<td>23</td>
<td>1,2,3</td>
<td>12:48</td>
<td>17</td>
<td>1.272727</td>
<td>1.000000</td>
<td>0.090909</td>
<td>3/3</td>
</tr>
<tr>
<td>73</td>
<td>3</td>
<td>12:49</td>
<td>17</td>
<td>1.504762</td>
<td>1.000000</td>
<td>0.380952</td>
<td>8/8</td>
</tr>
<tr>
<td>157</td>
<td>1,2</td>
<td>12:45</td>
<td>19</td>
<td>1.260000</td>
<td>1.000000</td>
<td>0.075000</td>
<td>3/3</td>
</tr>
<tr>
<td>220</td>
<td>1,2</td>
<td>12:40</td>
<td>17</td>
<td>1.400000</td>
<td>1.000000</td>
<td>0.250000</td>
<td>5/5</td>
</tr>
<tr>
<td>225</td>
<td>1,2,3</td>
<td>12:41</td>
<td>19</td>
<td>1.328000</td>
<td>1.000000</td>
<td>0.160000</td>
<td>4/4</td>
</tr>
</tbody>
</table>

Table 2: Results obtained by Kraft et al.'s method in the Cranfield collection

<table>
<thead>
<tr>
<th>#q</th>
<th>Run</th>
<th>T</th>
<th>Sz</th>
<th>Fit</th>
<th>P</th>
<th>R</th>
<th>#rr/#rt</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2,3</td>
<td>13:04</td>
<td>19</td>
<td>1.393103</td>
<td>1.000000</td>
<td>0.241379</td>
<td>7/7</td>
</tr>
<tr>
<td>2</td>
<td>2,3</td>
<td>12:39</td>
<td>17</td>
<td>1.488000</td>
<td>1.000000</td>
<td>0.360000</td>
<td>9/9</td>
</tr>
<tr>
<td>23</td>
<td>1,2</td>
<td>13:02</td>
<td>19</td>
<td>1.369697</td>
<td>1.000000</td>
<td>0.212121</td>
<td>7/7</td>
</tr>
<tr>
<td>73</td>
<td>1,2,3</td>
<td>12:38</td>
<td>19</td>
<td>1.619048</td>
<td>1.000000</td>
<td>0.523810</td>
<td>11/11</td>
</tr>
<tr>
<td>157</td>
<td>1,3</td>
<td>13:28</td>
<td>19</td>
<td>1.380000</td>
<td>1.000000</td>
<td>0.225000</td>
<td>9/9</td>
</tr>
<tr>
<td>220</td>
<td>3</td>
<td>13:40</td>
<td>19</td>
<td>1.640000</td>
<td>1.000000</td>
<td>0.550000</td>
<td>11/11</td>
</tr>
<tr>
<td>225</td>
<td>1</td>
<td>13:18</td>
<td>19</td>
<td>1.520000</td>
<td>1.000000</td>
<td>0.400000</td>
<td>10/10</td>
</tr>
</tbody>
</table>

Table 3: Results obtained by our SA-P IQBE method in the Cranfield collection

As regards the computation time required, both methods take approximately the same, although the SA-P seems to be a little bit slower. We think that this is a consequence of the inclusion of a selection criterion to get simpler queries in Kraft's method (see Section 3) which is not considered in the SA-P. This criterion makes the population being composed of simpler queries while the EA is converging, thus making their evaluation—the more time consuming procedure in both algorithms—less demanding.

6. Concluding Remarks

A new IQBE for FIRSs based on a hybrid SA-GP algorithm is been proposed and tested against the well known Kraft et al.'s proposal, outperforming the latter in the 1400 document Cranfield collection in terms of retrieval performance.

An extension of our method allowing it to adapt the retrieval threshold, which is usually a fixed value provided by the user, is also proposed in (Cordon et al., 2002). This increases has also been the system effectiveness even more without augmenting the algorithm run time at all.
Notes
1. Notice that the composition of several components is not the original one proposed by Kraft et al. but they have been changed to improve the algorithm performance. Of course, the basis of the algorithm have been maintained.
2. SA (Aarts, 1989) is a neighborhood search algorithm which modifies the usual acceptance criteria of the basic local search sometimes permitting accepting a worse solution than the current one to avoid getting trapped in local optima. SA starts from an initial solution and then generates a new candidate solution (close to it) by applying random changes on it. If the candidate solution is better than the current one, then the former replaces the latter. Otherwise, it still could be randomly accepted with a probability that depends on the difference between both solutions and on a parameter called temperature. This temperature is initialized to a high value (meaning that significantly worse candidate solutions are likely to be accepted) and then this value is decreased by a procedure called cooling strategy each time a number of neighbors is generated.

References

Abstract: The representation of information contents by graphical maps is an extended ongoing research topic. In this paper we introduce the application of UDC codes for the subject maps development. We use the following graphic representation methodologies: 1) Multidimensional scaling (MDS), 2) Cluster analysis, 3) Neural networks (Self Organizing Map – SOM). Finally, we conclude about the application viability of every kind of map.

1. Introduction

Advanced techniques for Information Retrieval (IR) currently make up one of the most active areas for research in the field of library and information science. New models representing document content are replacing the classic systems in which the search terms supplied by the user were compared against the indexing terms existing in the inverted files of a database.

One of the topics most often studied in the last years is bibliographic browsing, a good complement to querying strategies. Since the 80’s, many authors have treated this topic. For example, Ellis establishes that browsing is based on three different types of tasks: identification, familiarization and differentiation (Ellis, 1989). On the other hand, Cove indicates three different browsing types: searching browsing, general purpose browsing and serendipity browsing (Cove, 1988). Marcia Bates presents six different types (Bates, 1989), although the classification of Bawden is the one that really interests us: 1) similarity comparison, 2) structure driven, 3) global vision (Bawden, 1993). The global vision browsing implies the use of graphic representations, which we will call map displays, that allow the user to get a global idea of the nature and structure of the information in the database.

In the 90’s, several authors worked on this research line, developing different types of maps. One of the most active was Xia Lin what introduced the concept of Graphical Table of Contents (GTOC), comparing the maps to true table of contents based on graphic representations (Lin 1996). Lin applies the algorithm SOM to his own personal bibliography, analyzed in function of the words of the title and abstract fields, and represented in a two-dimensional map (Lin 1997). Later on, Lin applied this type of maps to create websites GTOCs, through a Java application.

The vectorization by words of the title and abstract fields (co-word analysis), generates too large a matrix, but this technique is applied to few documents. In this context, it is important to find some element that allows a lighter vectorization. The online library catalogs (OPACs) have diverse elements that can be vectorized with more ease than the free text, for example, the subject codes of Universal Decimal Classification (UDC). The objective of this work consists of verifying if it is possible
to create map displays using the UDC codes (co-classification analysis), for the purpose of creating GTOCs of the collection of a library.

2. Materials and methods

The selected OPAC for the study is that of the Public Library of Granada catalog which contains 32,700 records and 43,900 UDC codes, an average of 1.34 codes per record. These records were vectorized by the UDC codes, in such way that they are grouped in 27 major subject categories. The hierarchical structure that the UDC dominates taking advantage. To this matrix of data (17 x 32700), we applied the Pearson correlation index to measure the similarity among each one of the 27 major classes and to generate a new matrix (17 x 17), to which the methodology of space reduction will be applied.

For the creation of the display maps we opted for two major headings: 1) those of the statistics approach (based on multivariate analysis); and 2) those of the connectionist approach (usually, but not exclusively, based on artificial neural networks (ANN)).

Within the techniques of multivariate statistical analysis, we find three basic methods: 1) cluster analysis, 2) principal component analysis (PCA), and 3) multidimensional scaling (MDS) (Egghe, 1990). According to Kinnucan (Kinnucan, 1987), “These methods are referred to as dimensionality-reduction methods because this function is to simplify what might at first appear to be a complex pattern of associations among many entities.”

We described them in a previous study (Moya, 1998b) as follows:

1. Cluster analysis. This technique is used to create a two-dimensional display (e.g. dendrograms) of clusters of different objects whose relationships are represented by matrix values. This type of automatic classification, also known as numerical taxonomy, currently comprises more than 150 different techniques that are grouped in families according to shared procedures. In general, the IS discipline involves polythetic clustering hierarchies to create trees illustrating the hierarchy of relationships among elements on the basis of individual characteristics.

2. Principal component analysis (PCA). The basic premise of PCA is that the linear relation between any two variables is best summarized by a regression line. In other words, the variable that represents the regression line as a point cloud contains essential information about both variables. The two variables are thus combined into a single factor. This mechanism can be used to reduce pairs of variables to single dimensions in order to simplify the graphic display of the elements included in the matrix.

Multidimensional scaling (MDS). This multivariate analysis technique is used to identify the dimensions that best explain similarities and differences between variables. Because the purpose of MDS is to generate a map of objects, this approach can be considered an alternative to PCA.

Meanwhile, neural networks can learn to assign multidimensional outputs to multidimensional inputs, and they do so while maintaining a great capacity for generalization. We use the Self Organizing Map algorithm, also called Kohonen’s Feature Map. Kohonen’s interest in discovering how an organization of this type might arise led him to investigate the subject (Kohonen, 1982;1997). The product of
those researches was the network model that bears his name, and which is capable of performing a topological organization of the inputs presented to it.

This type of network has recently been used in documentation for domain analysis (White, 1998), for textual data mining (Kaski, 1998), to extract semantic relationships between words from their contexts (Honkela, 1995), and in particular to generate topological maps of sets of documents, even labeling the zones of influence of each word or term (Kaski, 1998; Moya, 1998a; 1999; Chen, 1998; Orwig, 1997; Guerrero, 2001).

3. MDS maps

In the MDS based display map (figure 1) we can see how each major subject category is placed in a certain point, in function of the relationships that borne to each other. Also, each category is represented with a proportional circle to the volume of documents that it contains. The circles appear in the center of the map, while the biggest are in the periphery, following the principle of center/periphery, established by White and Griffith (White, 1981)

We can also see that the categories are classified in two big clusters: 1) Science & Technology, and 2) Social Science & Humanities. This classification is corroborated by means of clustering, just as seen in the dendograma of figure 2. There are only two categories that don't seem to be in the recommended cluster (Economy and Law): however, we should keep in mind that the Economy is related with categories of the cluster Sci & Tech (Mathematics). On the other hand, the MDS places both categories at the edge of the map; in this way the dividing line can integrate them together with the Social Sciences area.

4. SOM maps

The map display based in SOM is quiet different than the MDS map (figure 3). The SOM map is more schematic, ordered and clear-to-see. Contrary to the MDS, the size of each category is not proportional to the volume of documents and this can give a confused view for the user. It is also very difficult to see the classification in two big clusters like the previous case. The categories group in function of their neighborhood relationships, to typical Kohonen’s algorithm feature, and it is lost from view the general structure that one could observe in the previous case.

The neighborhood relationships among the categories indicate the frequency of the co-occurrences of the classification codes. It is important to point out that the SOM looks for the good topology. This implies that when having to reduce to two dimensions the representation, the areas spread and they occupy their place in function of the greater or lesser contact among them. The proximity/distance among the areas is conditioned by co-classification frequency, however, which doesn't mean that the codes of classification of two far away categories to each other do not co-occur in an absolute way.

With regard to the forms of the areas, they are also determined by the co-classifications. These relationships cannot always be represented by regular areas, for which reason they take on more or less original forms.
5. Conclusions

Despite the fact that user-based evaluation experience of this kind of map display is very limited, we can assume the following conclusions:

- MDS and SOM are algorithms that can be used to generate bibliographic map displays.
- An OPAC can be represented through co-classification analysis, using UDC codes.
- MDS maps are better to view the structure of relations among the subject categories.
- SOM maps are easy-to-use, because the view is clear, schematic and ordered.
- SOM is easier to compute than the MDS, especially with a lot of variables.

References


![Figure 1 - MDS Map](image-url)
Figure 2 - Cluster dendogram

Figure 3 - SOM Map
Managing Documents with Bayesian Belief Networks: A Brief Survey of Applications and Models

Abstract: Belief networks are very appropriate tools to deal with uncertainty. They have been successfully applied to different environments. Information Retrieval has been one of these fields, where this property (uncertainty) is a very important feature, which has not been well managed by the classical models. In this paper, we present a brief overview of the different applications found in the specialised literature where these graphical models have been used to solve specific problems or to design complete Information Retrieval Models.

1. Introduction

The entire Information Retrieval (IR) cycle may be classified as an uncertain process because (Turtle & Croft, 1997): (1) The construction of the representations of documents and queries gives as a result incomplete characterisations, in the form of a set of terms. (2) The submitted query is just a vague description of the users' information need. (3) The computation of the relevance degree of each document with respect to the query inherits the previous uncertainty sources, but also presents the problems that can cause the different representations that a concept may have, as well as, that these concepts are not independent among them.

Probabilistic models (Crestani et al., 1998) tried to solve these problems although they have some limitations to overcome them. The development of Belief Networks and their applications to actual problems, with good results, caused that several researchers in the field of IR focused their attention on them. They realised that these kind of networks could be adequate models to be employed in IR, specially designed to work with a high performance in environments in which uncertainty is a very important feature, as the case of IR is. Also, because they can properly represent the relationships among variables.

The precursors of Belief Networks on IR were the studies carried out on Dependence Trees in the Probabilistic Model context. They captured the strongest dependences among a set of variables (in this case, terms) by means of Maximum Weight Spanning Trees (MWST)\(^1\), automatically constructed using the Chow-Liu algorithm (Chow & Liu, 1968) starting from term co-occurrences. These MWSTs
were initially utilised to evaluate documents and expand queries (van Rijsbergen, 1971; van Rijsbergen et al., 1981).

Later, some other applications were developed based on Belief Networks. In this paper we are going to briefly present some of them. To achieve this aim, this paper is divided into the following sections: After a brief introduction to the concept of Belief Network, Sections 3 and 4 will show the applications on Indexing and Hypertext, respectively. Later, we will describe the most important IR models developed with these tools. Section 6 deals with the feedback methods designed for these models. Finally, we end this paper with a section showing some concluding remarks.

2. What is a Belief network?

Belief Networks are graphical models able to represent and efficiently manipulate n-dimensional probability distributions (Pearl, 1988). A Belief Network uses two components to codify qualitative and quantitative knowledge, respectively: (a) A Directed Acyclic Graph (DAG), $G=(V,E)$, where the nodes in $V$ represent the random variables from the problem we want to solve, and the topology of the graph (the arcs in $E$) encodes conditional (in)dependence relationships among the variables (by means of the presence or absence of direct connections between pairs of variables); (b) a set of conditional distributions drawn from the graph structure. When these distributions are expressed by means of probabilities, then the Belief Network is called Bayesian Network. Therefore, for each variable $X_i \in V$ we have a family of conditional probability distributions $P(X_i | Pa(X_i))$, where $Pa(X_i)$ represents any combination of the values of the variables in $Pa(X_i)$, and $Pa(X_i)$ is the parent set of $X_i$ in $G$. From these conditional distributions we can recover the joint distribution over $V$:

\[ P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i | Pa(X_i)) \]  

This decomposition of the joint distribution gives rise to important savings in storage requirements. It also allows, in many cases, to efficiently perform probabilistic inference (propagation), i.e., to compute the posterior probability for any variable given some evidence about the values of other variables in the graph.

Once we know what a BN is, we are going to briefly focus our attention on its construction. There are two main approaches: (1) To manually build the network (using the knowledge provided by an expert, for instance. (2) Starting from a set of data, to apply an automatic process (learning algorithm) that obtains the most important (in)dependence relationships among the different variables that are represented on a DAG. Also the assessment of probability distributions could be manual or automatically done.

Finally, the BN is ready to carry out the inference process: Given a set of evidences, i.e., several variables whose values they take are known, they are instantiated on the network. By means of a probability propagation method, the probability that each variable takes a certain value given the evidences, is computed. If we take into account that the learning of a general BN as well as the propagation could be very time consuming tasks, the topology of the network is sometimes restricted, in the sense of reducing the complexity of the underlying
graph, and therefore, losing a bit of precision in terms of the (in)dependences that
the structure could represent. This action may reduce learning and propagation
times.

3. Applying Bayesian Networks to Indexing

In (Tzeras & Hartman, 1993), the authors have developed an indexing
technique based on the construction of a BN per document. It helps to decide the set
of descriptors, from a controlled vocabulary, that will be assigned to each
document, starting from a set of terms (single words or phrases) directly obtained
from the text. The network, composed of binary variables, is manually constructed
linking in the graph, in a first step, the nodes representing the terms found in the
text with the classes (referred by the authors as Forms of Occurrence, FOC) to
which they belong to. The links will be directed from the latter to the former. The
second step will join, using a rule dictionary, the terms with their associated
descriptors (directing the links from terms to descriptors). Finally, a node
representing the document is inserted as root and connected to the FOCs nodes.
Once the probability distributions have been assessed, the inference is carried out
by instantiating the document node and obtaining the belief on each descriptor
node, selecting those descriptors that have the highest posterior probabilities.

4. Applying Bayesian Networks to Hypertext

In the Hypertext environment, the two main studies that employs BNs (Frisse and
Cousin, 1989; Savoy and Desbois, 1992) work with a tree as the structure to organise the
index space, used to carry out tasks like query expansion or searching points to put into
practice a spreading activation process. On the one hand, in the first work, Frisse and Cousin
manually construct the graph, composed of medical terms. On the other hand, Savoy and
Desbois use the Chow-Liu's algorithm to automatically learn the tree. With respect to
probability distribution estimation, the latter works with an adaptation of the Jaccard's
coefficient (van Rijsbergen, 1971). The former assesses them using empirical methods. Both
approaches run the Pearl's propagation algorithm (Pearl, 1988) to compute the belief of the
terms.

5. Information Retrieval Models based on Bayesian Networks

The first complete IR model was the Inference Network Model (Turtle and
Croft, 1991), which is composed, in its simplified form, by two networks: the
document and query networks. The former represents the document collection and
contains two kinds of nodes: the document nodes, representing the documents, and
the concept nodes, symbolising the index terms contained in the documents. The
arcs go from each document node to each concept node used to index it. The
document network is fixed for a given collection. However, the query network is
dynamic, in the sense that it is specific for each query, and is composed by three
types of nodes: The Information Need node (inn), which represents the user's
generic information need; a set of intermediate query nodes, used in case of having
multiple query representations, and, finally, the query concept nodes (in the
simplified form, they are just the concept nodes in the document network, and
represent the connection between the two networks). The arcs in the query network
go from query concept nodes to query nodes and from query nodes to the
Information Need node. Each type of node stores a probability matrix, called link matrix in their notation, that in certain cases, depends on the type of query being formulated (Boolean or probabilistic). The retrieval is done by instantiating a single document node $D_j$ each time, and computing the probability that the Information Need is satisfied given that this document has been observed, $p(\text{inn} | d_j)$. Actually, Turtle and Croft precompute the intermediate probabilities $p(t_j | d_j)$ in the document network, and, later, use closed-form expressions to evaluate $p(\text{inn} | d_j)$ as a function of the probabilities $p(t_j | d_j)$, for those terms $T_j$ in the query.

The Ribeiro and Reis' model (Reis, 2000) is designed to simulate the Vector Space, Boolean and Probabilistic models. Their network is composed of three types of nodes: document nodes, concept nodes, and the query node. The arcs go from concept nodes to the document nodes where they occur, and from the concept nodes (appearing in the query) to the query node. In this model, the probabilities of interest are $p(d_j | Q)$, which could be computed as where $\tau$ represents any of the $2^M$ assignments of values to all the terms in the collection. This computation is obviously unfeasible. So, depending on the model to be simulated, the probabilities $p(Q | \tau)$ and $p(\tau)$ are defined in such a way that all the terms in the previous addition except one (corresponding to a given configuration $\tau_0$) are always equal to zero. Thus, the computation in equation (2) becomes straightforward: the inference is reduced to evaluate a function $(p(d_j | \tau_0))$ in the only non-zero configuration.

A third main IR model is called Bayesian Network Retrieval Model (BNRM) (de Campos et al., 2000; Fernández-Luna, 2001). It is composed of two sets of binary variables (relevant or non-relevant) organised in two subnetworks: The one containing the nodes that represent the documents of a given collection, and the network storing the nodes related to the terms by which the documents have been indexed. The subnetworks are linked between them: there is a directed link from the term node to the node associated with the document that it belongs to. The document subnetwork is initially organised with all the nodes isolated among them, expressing that documents are conditionally independent given the terms they contain (Also, it may include document relationships without any modification of the initial topology (de Campos et al., 2002)). The term subnetwork may have different topologies: A flat one, where the terms are independent among them (simple network), or an extended one, in which dependence relationships are allowed by means of a polytree (extended network). The reason is that using a more general structure could cause problems in the learning and propagation processes, due to the great number of terms that a collection may have. For polytrees, there are efficient algorithms that make possible the use of these structures in IR.

Once the topology is completely specified, we have to estimate the probability distributions stored in each node: nodes without parents will store a marginal distribution. The rest of nodes, a conditional probability distribution given the set of parents. In the case of document nodes, and due to the great amount of terms that could index a document, the probability matrices could have a very big size, producing important problems in the estimation time and storage. The solution
is the use of a probability function, that is called when the propagation algorithm needs a conditional probability, computing the required probability, and therefore avoiding to estimate all the conditional probability matrices.

Given a query submitted to the IR system by an user, the terms are instantiated to relevant and a propagation algorithm is run to compute the probability of relevance of each document given that the terms of the query are relevant.

Even taking into account the simplest topology that this network may have, the use of general inference algorithms in the whole network can not be applied due to efficiency considerations. This is the reason why we designed an approximate propagation technique, propagation + evaluation, exact when the probability function fulfils certain conditions. This method consists on an exact propagation on the term subnetwork (the two possible topologies allows an efficient application of a propagation algorithm), obtaining the probability of relevance of each term given the query terms, and later, evaluating a slightly modified probability function in the document nodes, to finally compute the belief of each document.

A first difference of this last model with respect to the Inference Network Model is that the first does not have a query network. A second distinction, also topological, is that the arcs in the BNRM are directed in the opposite way (from term nodes to document nodes). The BNRM authors think that is more intuitive to speak about the probability that a document is relevant given a query than the opposite. Therefore, their choice implies to instantiate the query and propagate towards document nodes. This fact means that in the BNRM only one propagation has to be carried out, unlike Turtle and Croft's model in which they have to run one propagation per document. With respect to propagation, the inference method in the simple BNRM, the most similar model to INM, allows to propagate in the whole network by only estimating prior probabilities and evaluating probability functions.

The network topologies of the Belief Network Model and BNRM are very similar, except by the fact that the last does not consider a query node. The main differences appear in the conditional probability distributions considered, in the BNRM case these distributions are not "degenerated" and do not depend on the query, and we truly perform probability propagation. Another important difference between these first two models and BNRM is that the last includes direct relationships between terms, thus obtaining a more expressive model.

The reader may study other two IR models based on BN in (Ghazfan et al., 1996; Bruza and van de Gaag, 1994).

6. Bayesian Network-based Information Retrieval Models and Relevance Feedback

Like in other IR models, those based on BNs have been also extended to include Relevance Feedback (RF) methods, improving in this way their performance with the user's help.

Let us begin with the RF designed for the Inference Network Model proposed in (Haines and Croft, 1993), which is carried out in two steps: (1) Query re-weighting: Modifying the probability distributions stored in the graph nodes, taking into account the information given by the judged documents. (2) Query expansion: The structure of the graph is changed by joining the query node to those new terms by which the expansion is done.
In the Belief Network Model (Reis, 2000), the network topology is also modified, inserting new document nodes corresponding to those documents that have been judged relevant. All these document nodes and the old query node are linked to a new query node. Later, the inserted document nodes are linked to the term nodes that represent the terms that index these texts. Finally, the probability distributions from new nodes have to be estimated.

The main difference between these previous RF methods is that while in the second technique, new terms are included in the network (and subsequently, the corresponding arcs), in the first one, only are added new arcs.

A completely different query modification approach is adopted by the Bayesian Network Retrieval Model (de Campos et al., 2001a, de Campos et al., 2001b). It is based on the concept of partial evidence and the way in which the Pearl’s propagation algorithm works. When a user evaluates the ranking returned by an IR system, he/she obtains new evidences about the relevance of the terms that belong to the judged documents with respect to the query. These evidences are quantified and included in the BN in propagation time. This approach presents the main advantage that there is no modification of the network structure, as well as any new computation of probability distributions.

7. Concluding remarks

In this paper we have briefly shown different applications and models of BNs in modern IR. We have seen how, due to their great expressiveness and also good conditions to represent the knowledge in uncertainty environments, they have been successfully applied to several fields of IR as alternatives to other techniques or to solve specific problems.

The performance of the BN models could be classified as very good compared to other models (modern models based on other Artificial Intelligence techniques or classic models).

The research on IR using these tools is still in effect, promising good results. For instance, in this survey, we have not talked about Cross Lingual Information Retrieval (CLIR), field that is clearly gaining importance in the IR community. Although there are some studies about CLIR and BNs, the main work is still to be done because the characteristics of this problem are especially appropriate to be treated with these tools. A second problem is Document Classification, where Bayesian Classifiers have been used (Sahami, 1998), and BN may contribute with very good solutions.

Notes
1 A tree obtained from any graph, in which the sum of the weights of its links is maximum.
2 From this point and for the rest of the paper, we will use the term “Bayesian Network” and its acronym BN.
3 Graphs in which there is no more than one directed path connecting each pair of nodes.

References


The Methodology of Constructing Classification Schemes: A Discussion of the State-of-the-Art

Abstract: Special classifications have been somewhat neglected in KO compared to general classifications. The methodology of constructing special classifications is important, however, also for the methodology of constructing general classification schemes. The methodology of constructing special classifications can be regarded as one among about a dozen approaches to domain analysis. The methodology of (special) classification in LIS has been dominated by the rationalistic facet-analytic tradition, which, however, neglects the question of the empirical basis of classification. The empirical basis is much better grasped by, for example, bibliometric methods. Even the combination of rational and empirical methods is insufficient. This presentation will provide evidence for the necessity of historical and pragmatic methods for the methodology of classification and will point to the necessity of analyzing “paradigms”. The presentation covers the methods of constructing classifications from Ranganathan to the design of ontologies in computer science and further to the recent “paradigm shift” in classification research.

1. Introduction
Classification of a subject field is one among about eleven approaches to analyzing a domain that are specific for information science and in my opinion define the special competencies of information specialists (Hjørland, 2002a). Classification and knowledge organization are commonly regarded as core qualifications of librarians and information specialists. Seen from this perspective one expects a firm methodological basis for the field. This paper tries to explore the state-of-the-art concerning the methodology of classification.

2. Classification: Science or non-science?
As it is part of the curriculum at universities and subject in scientific journals and conferences like ISKO, one expects classification/knowledge organization to be a scientific or scholarly activity and a scientific field. However, very often when information specialists classify or index documents and when they revise classification system, the methods seem to be rather ad hoc. Research libraries or scientific databases may employ people with adequate subject knowledge. When information scientists construct or evaluate systems, they very often elicit the knowledge from “experts” (Hjørland, 2002b, p. 260). Mostly no specific arguments are provided for the specific decisions in these processes.

If classification is just a question of adequate subject knowledge, then obviously any claim that classification should form part of a profession or a science is highly problematic (what many people actually find it/us to be).
There is at least one more way in which classification can be considered a non-scientific activity. Classification schemes may be seen as standards, reflecting compromises between different ways to look at fields. People working on classifications may have relevant subject qualifications, they may consult different sources and on that basis they may put some proposals. Those proposals may be changed or accepted as "standard". Such a standardization activity is not a scientific activity because it is not an explicitly argued activity in which published arguments are defended against other published arguments.

As researchers and teachers at universities we are, in my opinion, obliged to use research-based methods for the construction and evaluation of classifications, and that is why this paper explores the basis of our field.

3. Traditional methodologies for classification in LIS

The most established methodology for classification in LIS is probably the facet-analytic tradition founded by Ranganathan and further developed by the Classification Research Group (e.g. Mills, 1957 and Vickery, 1960). The second edition of the Bliss Bibliographical Classification has a well-described methodology (Mills & Broughton, 1977-), which has been applied in many different disciplines. According to Miksa (1998) the influence of this approach seems also to be increasing in the Dewey Decimal Classification.

Although I do find this approach an important part of our professional heritage, there are some obvious problems in the research tradition:

a) Its lack of explicit empirical methods
b) Its lack of methodological updating
c) Its lack of comparing its own approach with other approaches.
d) Its lack of formal recognition within LIS

Concerning a). Although the Bliss system must be praised for its rather detailed and explicit methodology, it is remarkable that no empirical methods are described. The construction of the classification is obviously built on some collection(s) of subject literature and scientific terminology, but no explicit methodological criteria are provided for empirical problems in this tradition.

Concerning b). The printed methodological works in the facet analytic traditions are rather old. The same principles are followed in each new volume of the Bliss system, and the concrete construction of all the disciplinary volumes is an impressive effort. From the perspective of classification as a scientifically based activity it is, however, a sign of crisis that no continuous methodological improvements are being suggested in the scientific journals.

Concerning c). Miksa (1998) pointed out that neither Ranganathan himself nor his followers have ever critically analyzed the basis of his approach. They have also paid insufficient attention to the strong points of other systems or to the weak points of their own. This is also a sign of a missing scientific basis.

Concerning d). No classification researchers are visible in bibliometric maps of LIS such as that provided by White and McCain (1998). This is one indication of the lack of formal recognition of classification as a proper field of research, even in our own discipline. This problem may be related to what research methods are taught in schools of LIS. Standard behavioral methods (whether quantitative or qualitative) are much better suited for, for example, user studies. In spite of problems in user studies (see Hjørland 2002a), such studies are prevailing probably
because they are much easier to conduct and much better suited for standard social scientific research methods. The teaching of research methods should, however, accommodate to the research problems and not the other way round.

Approaches such as algorithms for information retrieval (IR) and bibliometrics are not normally considered methods of classification. In my opinion they should be. They both assign elements of some kind to classes (or clusters, taxons, categories etc.), and they are both used for IR. Algorithmic methods are rather pure empirist methodologies, and bibliometrics combine aspects of empirical and pragmatic methods because it clusters documents that have previously been used together. These methods are relatively traditional within information science today (although not regarded as classification research). Few attempts have been made to combine such different methodologies. Rees-Potter (1989, 1991) is one remarkable exception. She described a method for semi-automatically constructing and maintaining a thesaurus. The method uses citation and co-citation analysis and is based on the hypothesis that highly cited papers are concept symbols.

The experimental IR tradition has been rather skeptical about classification. In this tradition the implicit view has been that an algorithm may classify documents based on term frequencies and that this approach will make “classification” superfluous. But will it? Algorithms may indeed classify documents and may even produce hierarchies. What do traditional classifications contribute in addition to this? The fact is that traditional classification involves structures that cannot be produced by any empirical analysis of the documents (or of the users for that matter). A geographical structure, for example, places different regions in a structure that is autonomous in relation to the documents that are written about those regions. You cannot produce a geographical map of Spain by making, for example, bibliometric maps of the literature about Spain. In my experience the classification of books in such autonomous structures is often regarded as highly useful (e.g. the regional catalogue in the Royal Library in Copenhagen). This is a fact which seems to have been overlooked in experimental IR and bibliometric approaches. From the point of view of classification research it is, however, a problem that such autonomous structures as geographical maps are provided by other specialists (e.g. geographers). We cannot claim to be experts in producing geographical maps, biological taxonomies or related structures. We can, however, explore their utility for IR just as we can explore general questions related to the methodology and philosophy of classification.

4. Methodologies for building ontologies

It has been claimed that the building of ontologies in computer science can be viewed as a technological take-over of classification research in LIS (Soergel, 1999 and Vickery, 1997). The stages of constructing an ontology has been described by Goble, Stevens and Karp (2000):

Identifying purpose and scope:

Knowledge Acquisition

Building the ontology

Identifying the intended range of uses of the ontology
The process of acquiring domain knowledge from which the ontology will be built
Conceptualization: identifying the key concepts that exist in
Building the ontology
Integrating: using or specializing an existing ontology as a foundation
Building the ontology
Encoding: representing the conceptualization in some form of language
Documentation
Informal and formal complete definitions
Evaluation
Determining the appropriateness of an ontology for its intended application

The construction of ontologies faces, however, the same kind of basic methodological problems as classification: problems related to obtaining domain knowledge and to uncovering how certain proposals support some kind of activities while hindering other activities.

5. Some innovative tendencies in classification research

Beghtol (1998, p.8) wrote: “a paradigm shift in bibliographic classification research is needed and may be developing”. She also introduced the concept of “cultural warrant” as an extension of the classical concept of “literary warrant” in classification. Central in the emergent paradigm are tendencies towards more historical, cultural and social understandings of knowledge, its production, organization and use. Innovative tendencies in classification research are also connected to attempts to base classifications on pragmatic methods (analysis of goals, values and consequences) rather than just to base it on the “neutral” or “objective” methods of rationalism and empiricism. Classification is an inherently social practice and, as such, cannot be understood without reference to the larger forces in which it, as a social practice, is embedded. Star (1989) has proposed that the Turing test, which is intended to measure the degree to which an expert system is able to perform as a human expert in its interaction with individual users, should be replaced by a “Durkheim test,” where the system is evaluated on its ability to support the goals of a specific community of users. Bowker and Star (1998) presents examples of classification research based on this innovative approach to classification.

Among other examples I can mention are Hansson (1999) who explored the classification system for Swedish libraries applying a “critical-hermeneutical method”. He shows how in 1921 the design of this system was shaped by the political and cultural currencies. Wallerstein, 1996 (reviewed in Hjörland, 2000) demonstrates how the classification of the social sciences has depended on basic philosophical assumptions that are seen as problematic today and thus points to the need for a restructuring of the social sciences. Hjörland (1998) has explored the classification of psychology and has demonstrated how this field tends to see itself as a “natural kind” but is perhaps better understood as a “social construction”. This paper also presents and discusses how empiristic, rationalistic, historicist as well as...
pragmatic methods have been applied to the classification of both psychological phenomena and psychological literature.

It should be mentioned that all kinds of technology face the same problem concerning the objectivity of classification. There is no escape from discussing values and goals. About Internet search engines Introna and Nissenbaum (2000), for example, demonstrate that such engines are not neutral, "Why the Politics of Search Engines Matter". The tendency to regard problems of indexing, classification and information retrieval as primarily technological problems is wrong because it ignores the qualitative questions about what kind of information/documents are made relatively invisible. In itself the focus on technology alone is an ideology which implies that what is already most visible becomes more visible and what is marginalized will be further suppressed.

In general these innovative methods are more "humanistic", "scholarly" and "soft" and may look less specific compared to traditional approaches. It is always possible, however, to draw specific conclusions about methodological principles when consequences related to different goals and interests have been uncovered.

6. Constructing classifications imply uncovering "paradigms"

In this section I shall propose a principle of classification based on Activity Theory (AT) as described in Hjorland (2002c). According to AT the concept of tools includes ideas, documents, classifications and symbolic systems. Tools have functional values in relation to certain cultures (or subcultures). Meanings, signs, documents and classifications have developed functionally to answer the needs of relatively stable forms of practices in human societies. For example, we use the bible and the hymnbook in our relatively stable religious practice, we use bodies of laws in our relatively stable legal practice and we use textbooks in our relatively stable teaching practice.

Concepts as well as documents have more or less stable meanings in relation to such applications, and there are built-in relations between concepts. As the anthropologists have demonstrated, different cultures tend to classify the world differently. Given classifications tend to fit more or less with the pre-understandings of given (sub)cultures. In a given community or subculture there is always more or less consensus on whether the existing practice is adequate or whether it should be changed. There will also be different opinions of whether the existing practice should be changed in one or another direction. One could say that in a given (sub)culture there always exist different "paradigms" for how practices should be changed. Such "paradigms" are related to different worldviews and interests etc.

Any change in practices implies a need to change the documents, the symbolic systems, the concepts and the classifications that support the existing practice.

Given concepts, documents and classifications will always serve certain policies and practices better than other concepts, documents and classifications. This is the case whether or not people are aware of this relationship.
Some meanings, classifications etc. are simply better suited for certain "paradigms" than for other paradigms. A given tool, e.g. a hammer, may be used for different purposes: to nail something, as a toy, as ballast, as a murder weapon etc. Some purposes are more conventional than others and the most conventional practices are most influential in determining meanings, documents and classifications. Words and meanings are applied to perform linguistic acts which are very often aiming at extra-linguistic acts. Meanings may be more or less suited for supporting the intentions of the language user. (The thoughts we think depend on the concepts we have learned from other people. Therefore, we may think by means of foreign concepts that are not suited for our own purpose). Different "paradigms" and interests tend to develop their own systems of concepts and meaning (signs, symbols, documents, classifications etc.). This is the case with different scientific paradigms as well as with political ideologies, religious and cultural paradigms etc. If, for example, a person terms certain localities "tourist attractions" that person facilitates a certain application of the place. Different concepts (and conceptual structures or classifications) facilitate some practices and hinder other practices relatively. When a librarian uses a certain concept or word about a document, she facilitates a certain application of the document at the expense of other applications.

Attempts to innovate something imply an attempt to change concepts, meanings and classifications in a way that makes them facilitate the desired innovation. Different views on a certain practice (e.g. a psychiatric practice) are associated with disciplinary theories and paradigms. Each paradigm tends to develop its own language, system of categories, documents and organization of knowledge. (E.g. psychoanalytic, neuroscientific or existential). A given word has often many different meanings associated with different practices and interests. The user of the word may not be aware of this ambiguity before he faces a problem and begins to examine the meaning of the word in different sources. Theoretical clarity is normally a relatively late stadium in the development of a paradigm. In a developed paradigm (e.g. Newton's mechanics) concepts are well defined, but not one by one. A system of mutually dependent categories and concepts are developed. The process of defining a concept (e.g. information, document, subject or concept) implies an attempt to make the concepts functional in relation to the theoretical work one is doing (e.g. adapting the discipline to the digital revolution).

6. Conclusion

To design a classification scheme (or to build an ontology) is a complex task involving the understanding of underlying concepts, conceptions, meanings and how those meanings support or hinder different kind of practices. Perhaps the most important step should be to uncover the different interests and "paradigms" in the domain. This can be done by using epistemological, historical, bibliometrical and other methods of domain analysis (cf. Hjørland, 2002a).

References


Abstract: This paper describes the cultural construction of classification as exemplified by the French Encyclopédists, Jean d'Alembert and Denis Diderot, and the encyclopaedism of Samuel Taylor Coleridge analysing original texts digitized and encoded using XML and an adaptation of TEI.

1. Introduction

This paper, focusing on encyclopaedism, is part of a larger study exploring the cultural construction of classification. The larger study explores possible foundations for bias in the structure of classifications with a view to more equitable practice. Bias in classification has been documented relative to race, ethnicity, gender, religion, sexuality and other factors. Analyses and proposed solutions have addressed only acute biases in particular systems, not the systems themselves. The project tentatively identifies the systemic roots of bias are culturally specific and reflected in the structure of conventional classificatory practices. A wide range of western cultural texts from classic Greek philosophy to twentieth-century ethnography is being analysed. The consistency with which certain presumptions are revealed, no matter how different the philosophical and social views of the authors, indicates their ubiquity in western thought, though it is not mirrored in many other cultures. We hope that an understanding of these fundamental cultural presumptions will make space for development of alternative approaches to knowledge organization that can work alongside conventional methods.

This paper describes an example of the first phase of the project, which is a deconstruction developed from relevant texts. In the context of encyclopaedism the key texts used in this paper are Jean d’Alembert’s *Preliminary Discourse* to the *Encyclopédie*, selections from Denis Diderot’s contributions to the *Encyclopédie*, and Samuel Taylor Coleridge’s *Treatise on Method* and *Prospectus of the Encyclopaedia Metropolitana*. We are analysing these texts in digital form using Extensible Markup Language (XML) implemented via a document type definition (DTD) created for the purpose including elements of the Text Encoding Initiative (TEI). We will first explain the encoding methodology; then define the differences between the French Encyclopaedists and the English Coleridge; deconstruct these differences by allowing the commonalities between the texts to emerge; and, finally, examine their cultural specificity.
2. Encoding texts

Encoding texts for analysis involves three stages: acquire each text in digital form, establish the DTD, encode the texts. The first stage began by identifying texts available in digital form from sources such as IPL Online Texts Collection: http://www.ipl.org/reading/books/ and http://www.ipl.org/reading/books/other.html and Project Gutenberg: http://www.promo.net/pg/. Completeness and reliability of the e-texts were verified as some "full-text" versions were actually partial and even where full e-texts were located, proofreading and editorial issues of originating agencies emerged. Texts not on the Web needed scanning and optical character recognition (OCR). This required obtaining copyright permission where we wanted to use more of the texts than allowed under fair. Permission was denied only once, but responses are slow and some publishers failed to reply. Such texts must be analysed from print, allowing comparison of the two approaches and assessment of digital encoding in deconstruction. For scanned texts, tagged image files (.tifs) were created, proofread and edited using two OCR software programs for comparison. The resulting text still required considerable tidying-up and close reading to ensure that nothing was lost in the conversion process from OCR to text document.

The second stage involved developing the DTD, which is a sort of authority list for the codes used to mark both explicit and implicit manifestations of key research themes. Exploring the potential of XML is an important component of the project. An XML version of the TEI Lite DTD was developed containing the necessary elements for prose textual analysis. We created our own set of thematic tags compatible with the TEI Lite DTD, facilitating human reading and XML parsing by providing more explicit identification of themes. It is effective to date in maintaining a format readable in a standardized structure while meeting the needs of our particular project.

Finally, with digitized texts and DTD in hand, we have begun encoding texts. Further encoding and development of full cross-referencing and a database continue to allow analysis of the texts from multiple perspectives. The remainder of this paper is a product of our initial work on representative encyclopaedist texts.

3. Encyclopaedist rivalry

Highlighting the obvious differences between the French Encyclopaedists and Coleridge's English reaction to them sets the stage for looking at their startling similarities in relation to classification. Three areas are sufficient to demonstrate differences between these two schools: their purposes, their historical grounding, and Coleridge's views on the French Encyclopaedists.

3.1 Purpose in encyclopaedism

The Encyclopaedists led by Diderot and d'Alembert created their Encyclopédie as a voice of reason denouncing the French crown and church. Their idealistic mission was: "to collect all knowledge ... to present its general outlines ... that our children, by becoming more educated, may at the same time become more virtuous and happier ..." (Diderot, 1751, p 92). They saw the corruption of crown and church as antithetical to reason and to the project of an enlightened populace as Diderot noted in his Encyclopédie entry on Consecrated Bread: "We are astonished that there is so much misery around us; and I for one, when I see all our extravagance and folly, am even more astonished that there does not exist much
more" (1751, p. 79). Statements like Diderot’s in his entry for Political Authority — "No man has received from nature the right to command others" (1751, p. 185) — are indicative of the subversive intent of this work. Such comments are common in most of the contributions to the Encyclopédie, illustrating the goal of Diderot and his colleagues to establish a reign of enlightened reason in place of the dominance of the church and the divine right of kings.

Early in the following century, Coleridge proposed his Encyclopedia Metropolitana. His focus was on “the vivifying power of Method” as he described in his Treatise on Method (1818, p. 53), one of the few parts of his endeavour to be published. Coleridge deplored the “worse than inmethodical” attempts of his continental predecessors. Alluding to the Encyclopédie, Coleridge laments the lack of logic in the arrangement of such works, a shortcoming to be remedied by application of method with the primary purpose of being “to the rising education of the country at once a reservoir and a fountain.” (1849, p. 72). Coleridge describes how “the fountains of education may be poisoned ... (through the] insinuation of sceptical principles into Works of Science ...” (1818, p. 52) by the French. In a sermon, Coleridge accuses them:

it seems to have been about the middle of the last century, under the influence of Voltaire, D'Alembert, Diderot, say generally of the so-called Encyclopaedists, and alasi — of their crowned proselytes and disciples, Frederick, Joseph and Catherine, — that the human understanding, and this too in the narrowest form was tempted to throw off all show of reverence to the spiritual and even to the moral powers and impulses of the soul; usurping the name of reason, openly joined the banner of Anti-Christ, at once the pander and the prostitute of sensuality, and whether in the cabinet, laboratory, the dissecting-room or the brothel, alike busy in the schemes of vice and irreligion” (1938, p. 109)

Coleridge articulated the differences between his view and that of the French Encyclopaedists. He believed “that reason is the knowledge of the laws of the whole,” a higher faculty than understanding, which is merely the result of observation (1938, p. 106). He links reason to the divine so the French Encyclopaedists, as bearers of the banner of the Anti-Christ, cannot possibly be genuine followers of reason, but only practitioners of a shallow empirical understanding. Coleridge filtered their critiques of crown and church were through the intervening horrors of the French Revolution. They lived in entirely different political and religious times.

3.2 Historical grounding

The French Encyclopaedists, acolytes of enlightenment, found their historical grounding in the European Renaissance. D'Alembert, in his Preliminary Discourse describes the Middle Ages when classical texts were available only in limited and often inaccurate form:

The masterpieces that the ancients left us in almost all genres were forgotten for twelve centuries. The principles of the sciences and the arts were lost, because the beautiful and the true, which seem to show themselves everywhere to men, are hardly noticed unless men are already apprised of them. Not that these unfortunate times were less fertile than others in rare geniuses; Nature is always the same. But what could these great men do, scattered as they always are, from place to place, occupied with different purposes, and left to their solitary enlightenment with no cultivation of their abilities? (1751, p. 61)
D'Alembert views the reintroduction of classical texts to Europe that spawned the Renaissance as the dawn of enlightenment: “On emerging from barbarism, the human mind found itself in a sort of infancy. It was eager to accumulate ideas, but incapable at first of acquiring those of a higher order because of the kind of sluggishness in which the faculties of the soul had for so long a time been sunk” (1751, p. 63). He tracks European thought following the Renaissance thinker Francis Bacon’s three stages of human development: memory, reason and imagination. D'Alembert and his colleagues took Bacon as the source of their classification and the basis of their view of history and development. Like so many thinkers of the European 18th-century Enlightenment, they were grounded in classicism, especially as introduced by the Renaissance.

Coleridge, however, regretted the loss of medieval order. He valued the classification of liberal arts and sciences of the trivium (grammar, dialectic, rhetoric) and quadrivium (arithmetic, geometry, astronomy, music) leading to the crowning study of theology in medieval Scholastic education: “the SCIENCE of Theology was the root and trunk of the knowledges that civilized man, because it gave unity and the circulating sap of life to all other sciences, by virtue of which alone they could be contemplated as forming collectively the living tree of knowledge” (1938, p. 167). He explained in his prospectus and preface to the *Encyclopedia Metropolitana* that the liberal arts and sciences gave method to knowledge organization (1818, pp. 71-72). The trivium and quadrivium were the basis for definition of disciplines as developed in the medieval university by the Scholastics, medieval scholars focused on method. Coleridge’s own focus on method makes this link unsurprising. D'Alembert, on the other hand, said the Scholastics “mistook for the true philosophy of the ancients a barbarous tradition which disfigured it” (1751, p. 62).

Coleridge and the French Encyclopaedists seem leagues apart.

4. Classificatory commonality

However, both D'Alembert and Coleridge assume that a unified circle of knowledge is not only possible, it is desirable. D'Alembert writes characteristically about the ideal of a unified view of knowledge: “The mind acquires reflective knowledge by making use of direct knowledge, unifying and combining it” (1751, p. 6). Here unification is through logic: “[t]he discovery of truth, which is the aim of Logic, produces the art of transmitting it to others” (1751, p. 54). Logic is used to bring reflective knowledge, the fruits of empirical observation, to an understanding of a unified view. Coleridge is similar in his quest for a “universal Method, by which every step in our progress through the whole circle of Art and Science should be directed” (1818, pp. 2-3). This universal structure of knowledge presumes three attributes: mutually exclusive categories, teleological progression, and the primacy of hierarchy (Olson, 1999).

4.1 Mutual exclusivity

D'Alembert introduces, in the *Preliminary Discourse*, the concept of ‘impenetrability’ which is a variation on the idea of mutually exclusive categories: In our study of Nature ... we note that bodies have a large number of properties. However, in most cases they are so closely united in the same subject that, in order to study each of them more thoroughly, we are obliged to
consider them separately. Through this operation of our intelligence we soon discover properties which seem to belong to all bodies, such as the faculty of movement or of remaining at rest, and the faculty of communicating movement, which are the sources of the principal changes we observe in Nature. By examining these properties — above all the last one — with the aid of our own senses, we soon discover another property upon which all of these depend: impenetrability, which is to say, that specific force by virtue of which each body excludes all others from the place it occupies, so that when two bodies are put together as closely as possible, they can never occupy a space smaller than the one they filled separately. (1751, pp. 16-17)

Impenetrability carves up space, physical or intellectual, into mutually exclusive categories.

Coleridge lays the basis for mutually exclusive categories by differentiating between law and theory as the ‘uniting’ and ‘progressive’ powers of method. Law is, for Coleridge, “a rule to which the subjects of the Law must necessarily conform” (1818, p. 4). Law is the unifying factor. The progressive factor is “THEORY, that in which the existing forms and qualities of objects ... suggest a given arrangement of them to the Mind ...” (1818, p. 4). Law, like Coleridge’s view of Reason, is linked to the “absolute perfection conceivable only of GOD” (1818, p. 4). Theory is of a “second order of relations” (1818, p. 58). Mutually exclusive categories begin with pure sciences (based on Law and reason) versus mixed sciences (based on theory and the mind) (1818, p. 58).

D’Alembert and Coleridge, through impenetrability and the distinction between law and theory, each set up the incompatible differences that construct mutually exclusive categories.

4.2 Teleology

Both D’Alembert and Coleridge exhibit teleology in two senses: progress in human culture and the logical progression of topics in classification. D’Alembert states that efforts at self-preservation caused humanity “to make some progress along the path of knowledge” (1751, p. 15). He develops this through a hypothetical description of motives. Logic, the “science of reasoning, which is rightly considered the key to all our knowledge” was a product of this progress (1751, p. 30). Coleridge took a similar view: “As in the individual, so in the whole community of Mankind, our cogitations have an infancy of aimless activity; and a youth of education and advance towards order; and an opening manhood, of high hopes and expectations; and a settled, staid, and sober middle age, of ripened and deliberate judgment” (1818, p. 47).

The development of thought and knowledge leads logically for both d’Alembert and Coleridge to the appropriate order of classification. D’Alembert largely follows Bacon by adopting the three main branches of Memory/History, Reason/Philosophy and Imagination/Art. Within each of these branches, subdivisions are developed in a logical progression:

The general distribution of beings into spiritual and material provides a subdivision of the three general branches. History and Philosophy are occupied with each of these two kinds of beings, while imagination deals only with purely material beings... At the head of the spiritual beings is God, who necessarily holds the first rank ... Below that Supreme Being are the created spiritual beings whose existence is taught us by Revelation. Next comes man. Composed of two principles, he belongs by virtue of his soul to the spiritual
beings and by virtue of his body to the material world. And finally comes that vast universe which we call the corporeal world, or Nature. (1751, p. 52)

Coleridge works through a list of disciplines clearly representing the legacy of the Scholastics in his explanation of “the Arts and Sciences in their Philosophical harmony” (1818, p. 55) as he works methodically from grammar leading to logic leading to mathematics and so forth to “the great Cause of all” (1818, p. 55) “for at the head of all Pure Science stands Theology” (1818, p. 57).

4.3 Hierarchy

These teleological chains of topics lead seemingly ineluctably to hierarchy:

We have shown that this METHOD consists in placing one or more particular things or notions, in subordination, either to a preconceived universal Idea, or to some lower form of the latter; some class, order, genus, or species, each of which derives its intellectual significance, and scientific worth, from being an ascending step toward the universal; from being its representative, or temporary substitute. Without this master-thought, therefore, there can be no true Method; and according as the general conception more or less clearly manifests itself throughout all the particulars, as their connective and bond of unity: according as the light of the Idea is freely diffused through, and completely illumines, the aggregate mass, the Method is more or less perfect.” (Coleridge 1818, p. 54)

Hierarchy is the result of nesting subordinate topics under superior ones — class, order, genus, species — just as above d’Alembert pursued subdivision of Bacon’s three main classes by spiritual and material and then by rank similar to the Great Chain of Being extending from God to Nature. D’Alembert uses the image of a tree of knowledge with its branches constantly dividing into more specific topics as does Coleridge when speaking of “the subordination which necessarily arises among the different branches of Knowledge, according to the difference of those Ideas by which they are initiated and directed; for there is a gradation of Ideas, as of ranks in a well-ordered State, or of commands in a well-regulated army ...” (1818, p. 9).

The development of a hierarchy is not, however, simply a matter of division. It is also a function of reduction, a reversal of division and mirroring induction. The particular leaves and branches of the tree lead to the unity of the trunk:

... it is not at all by vague and arbitrary hypotheses that we can hope to know nature; it is by thoughtful study of phenomena, by the comparisons we make among them, by the art of reducing, as much as that may be possible, a large number of phenomena to a single one that can be regarded as their principle. Indeed, the more one reduces the number of principles of a science, the more one gives them scope, and since the object of a science is necessarily fixed, the principles applied to that object will be so much the more fertile as they are fewer in number. This reduction which, moreover, makes them easier to understand, constitutes the true "systematic spirit." (d’Alembert, 1751, p. 22)

Coleridge concurs when he says that the task of his Encyclopaedia Metropolitana follows “the principles of unity and compression” (1818, p. 54). An arbitrary arrangement, such as an alphabetical one, is one in which “the desired information is divided into innumerable fragments ... like a mirror broken on the ground presenting, instead of one, a thousand images, but none entire ...” (1818, p. 66). It is the focusing capabilities of hierarchical organization that compresses and reduces the fragments into the generalized truth of a law.
5. Illusory universality

In spite of their differences, d'Alembert and Coleridge base their encyclopaedic efforts to organize knowledge on the same three classificatory principles of mutually exclusive categories, teleology and hierarchy. In doing so they draw from the different traditions of the medieval Scholastics and the classical revival of the Renaissance. The similarities in fundamental presumptions indicate the potential for these three classificatory principles to be universal. However, there are other factors that suggest they may be culturally specific to western thought.

The most obvious indicators of cultural specificity are comments on other cultures. Coleridge is frank in such matters. He imagines “an unlettered African, or rude, but musing Indian, poring over an illuminated manuscript of the inspired volume; with vague, yet deep impression, that his fates and fortunes are, in some unknown manner, connected with its contents” (1818, p. 52). Coleridge goes on to suggest that this African or Indian might well sort out the concepts, but, because the result is “without soul or substance, a talisman of superstition” it will lead only to death or vanity (1818, p. 51). “But see, the friendly missionary arrives!” translating and explaining, with enlightenment as the result (1818, p. 52). For Coleridge’s the African or Indian cannot attain enlightenment, because Method is required for infinite progression (1818, p. 7).

D’Alembert is more subtle in his view of other cultures. He does not specifically state an intolerance. Diderot writes against intolerance in his Encyclopédie on that topic (1751, pp. 152-156). However, the implication is still there in expressions on progress. D’Alembert gives considerable space in the Preliminary Discourse to the development of rational thought. He explains how self-preservation naturally leads to ethics and ethics to reason. The implication of this progression is that any cultures that do not adopt the same logic are not fully developed.

Teleology, in the sense of progression towards a goal, suggests that there is some end toward which societies or cultures are progressing. Hierarchy validates both the notion of superiority and that of reduction of differences to some universal truth. Both teleology and hierarchy tend toward universality. Teleology, as developed by both d’Alembert and Coleridge, suggests that there is a universal goal: reason (or logic). Hierarchy, especially in its tendency to reductionism, focuses differences into some unity—some grand category of truth.

Homi Bhabha defines “cultural difference” as “the process of the enunciation of culture as ‘knowledgeable’, authoritative, adequate to the construction of systems of cultural identification” (1994, p. 34). The alternative is homogenization. Teleology and hierarchy, and their necessary precursor mutual exclusivity, suggest that universality is the goal as well as the method. The result is that cultures not characterized by these principles are not able to have knowledge, are not authoritative and lack identity. “The concept of cultural difference focuses on the problem of the ambivalence of cultural authority: the attempt to dominate in the name of a cultural supremacy which is itself produced only in the moment of differentiation.” (1994, p. 34) Implementation of supremacy found in the hierarchies created to focus truth down to great universalities instead of broken fragments, denies differentiation. These two variant strains of encyclopaedism grow from the same genetic sources. They are integral parts of western culture. Their differences are superficial so that ultimately they present a united identity. If the classificatory principles of mutually exclusive categories, teleology and hierarchy
are allowed to engross classificatory practice, the result is a homogeneity that
denies difference and identity. Drawing boundaries between topics, sequencing
them in a culturally defined stream (such as Bacon's three main classes) and
subsuming some under others has the potential to erase cultural identity, especially
for cultures that foster the lateral and/or cyclical integration of overlapping or
porous concepts (see Olson, 2000).

References
Constable, 1934.
Diderot, Denis. (1751-1772). Denis Diderot's The Encyclopedia: Selections. Ed. and tr.
Olson, Hope A. (2000). Reading "Primitive Classification" and misreading cultures: The
metaphysics of social and logical classification. In Clare Beghtol, Lynne C. Howarth and
Jian Qin  
School of Information Studies, Syracuse University, NY, USA

Evolving Paradigms of Knowledge Representation and Organization: A Comparative Study of Classification, XML/DTD, and Ontology

Abstract: The different points of views on knowledge representation and organization from various research communities reflect underlying philosophies and paradigms in these communities. This paper reviews differences and relations in knowledge representation and organization and generalizes four paradigms—integrative and disintegrative pragmatism and integrative and disintegrative epistemologism. Examples such as classification, XML schemas, and ontologies are compared based on how they specify concepts, build data models, and encode knowledge organization structures.

1. Introduction

Knowledge representation (KR) is a term that several research communities use to refer to somewhat different aspects of the same research area. The artificial intelligence (AI) community considers KR as simply "something to do with writing down, in some language or communications medium, descriptions or pictures that correspond in some salient way to the world or a state of the world" (Duce & Ringland, 1988, p. 3). It emphasizes the ways in which knowledge can be encoded in a computer program (Bench-Capon, 1990). For the library and information science (LIS) community, KR is literally the synonym of knowledge organization, i.e., KR is referred to as the process of organizing knowledge into classifications, thesauri, or subject heading lists. KR has another meaning in LIS: it "encompasses every type and method of indexing, abstracting, cataloguing, classification, records management, bibliography and the creation of textual or bibliographic databases for information retrieval" (Anderson, 1996, p. 336). Adding the social dimension to knowledge organization, Hjørland (1997) states that knowledge is a part of human activities and tied to the division of labor in society, which should be the primary organization of knowledge. Knowledge organization in LIS is secondary or derived, because knowledge is organized in learned institutions and publications. These different points of views on KR suggest that an essential difference in the understanding of KR between both AI and LIS lies in the source of representation—whether KR targets human activities or derivatives (knowledge produced) from human activities. This difference also decides their difference in purpose—in AI KR is mainly computer-application oriented or pragmatic and the result of representation is used to support decisions on human activities, while in LIS KR is conceptually oriented or abstract and the result of representation is used for access to derivatives from human activities.

Despite the essential difference, these different versions of KR share some common principles and methodologies. For example, AI's KR stresses adequacy and expressiveness, e.g., a scheme that represents a knowledge domain should be
sufficient to allow any fact of interest to be inferred. Similarly, LIS’s KR emphasizes the importance of representing the same phenomenon in different contexts such as in sociology, economics, psychology, history, and so forth. Both use some encoding language and format for representation. This paper discusses KR in such a general way as described by Duce and Ringland (1988), that is, from a structural and language point of view rather than a computational point of view. By using examples of various knowledge structures, this paper presents four paradigms prevailing in KR research and practices and compares three knowledge organization structures to demonstrate how these paradigms impact them.

2. Paradigms in Knowledge Representation and Organization

Paradigms symbolize meta-theoretical assumptions about the nature of the subject of study (Berrell and Morgan, 1979), or “universally recognized achievements that for a time provide model problems and solutions to a community of practitioners” (Kuhn, 1970). Hirschheim and Klein (1989) resolve the differences between Berrell and Morgan and Kuhn by pointing out that a paradigm consists of a “most fundamental set of assumptions adopted by a professional community that allows its members to share similar perceptions and engage in commonly shared practices.” Even though differences exist in KR practices, common approaches are used across research communities. These common practices include hierarchical organization of concepts and horizontal relations between them. Let us examine the following examples of representing the concept of anthrax.

The Antibiotic Guide represents the concept of anthrax using a problem-solving approach. It divides issues surrounding anthrax into problem-solving areas
such as diagnostic criteria, common pathogens, treatment regimens, and important points. Under each problem-solving area, more specific concepts and solutions are defined. In MeSH, the concept of anthrax is represented through two tree structures: the *Bacteria* and the *Bacterial Infections and Mycoses*. Each of them is a typical hierarchy that integrates into a system with different levels of knowledge about anthrax.

<table>
<thead>
<tr>
<th>Bacteria</th>
<th>Bacterial Infections and Mycoses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endospore-Forming Bacteria</td>
<td>Bacterial Infections</td>
</tr>
<tr>
<td>Gram-Positive Endospore-Forming Bacteria</td>
<td>Gram-Positive Bacterial Infections</td>
</tr>
<tr>
<td>Gram-Positive Endospore-Forming Rods</td>
<td>Bacillaceae Infections</td>
</tr>
<tr>
<td>Bacillaceae</td>
<td>Anthrax</td>
</tr>
<tr>
<td>Bacillus</td>
<td></td>
</tr>
<tr>
<td>Bacillus anthracis</td>
<td></td>
</tr>
</tbody>
</table>

Example 2. The Concept of Anthrax in the *Medical Subject Headings* (MeSH)

These two examples pose an important and interesting question: Are they fundamentally different in representing the knowledge or are the representations simply some variations of the same paradigm that originates from the same principles or philosophies? Obviously, the answer to this question may not be a straightforward "yes" or "no." Organizing knowledge in libraries has a long history of using an integrated approach. Think about hierarchical and faceted classifications. Both structures integrate human knowledge into a systematic arrangement, in which concepts and structures tend to be abstract and have an epistemological orientation. On the contrary, newer knowledge technologies such as XML schemas and ontologies take an opposite approach in representing knowledge, which disintegrate parts of knowledge into a problem-solving focused structure and are more pragmatic and application-oriented. If we put these approaches together with two intercepting spectra, we obtain four paradigms as shown in Figure 1. The integration paradigm is best summarized in the theory of "integrative levels" (Feibleman, 1954). The integrative levels, as Feibleman states, represent some uniformity in science as well as the physical world. The integrative levels theory views each level of the physical world as an organization of the level or levels below it plus one emergent quality. The integrative levels are cumulative upward and complexity of the levels also increases upward; the higher level depends upon
the lower, and the lower is directed by the higher. For an organization at any given level, its mechanism lies at the level below and its purpose at the level above. The MeSH example is a good demonstration of the integrative levels theory.

The disintegration paradigm takes an opposite approach. Rather than organizing knowledge into a vertical hierarchy, disintegrative representations focus on the concept and all aspects related to it, which Ingwersen calls it "polyrepresentation" (1994). Disintegrative representations define a concept and match the right solutions to problems related to this concept. Consequently, the representations become the conceptual model or framework for an application. In this situation, where the concept is located in the knowledge system is less important than what solution areas there are in relation to the concept. The second example above demonstrates such an underlying statement.

Another way of considering KR paradigms is as pragmatic versus epistemological. "Pragmatism stresses the instrumentality of human knowledge and concepts." (Hjørland, 1997, 97) It takes practical consequences as the criteria of knowledge and meaning. Epistemologism engages in abstract and epistemological representations and structures. The intercepting areas as shown in Figure 1 form four distinctive paradigms: integrative pragmatism for which Dewey Decimal Classification (DDC) and Universal Decimal Classification are representative; integrative epistemologism as reflected in Colon Classification; disintegrative pragmatism showing a trend in newer knowledge technologies; and disintegrative epistemologism as represented by ontologies. Due to space limitations, an in-depth discussion of these intercepting paradigms will have to be given in another paper. They nevertheless raise a number of important questions: Do the paradigms underlie practices in both LIS and AI communities? In what ways the two communities perceive the paradigms? How have the paradigms affected the KR research and practices? Although addressing these questions is beyond the scope of this paper, a comparison between some KR structures is provided below to show how different paradigms might have influenced the representation outcomes.

3. Comparison of Knowledge Structures

Table 1 lists similarities and dissimilarities between three knowledge structures organized by the ways in which concepts are specified, data modeled, and knowledge is encoded. The variations among them are largely decided by the purpose of each representation. Classification is commonly used in almost every field of human activities and the physical world, including newer representation structures such as XML Document Type Definition (DTD) or schemas and ontologies. Library classification such as DDC (including some thesauri that have a covert classification hierarchy through a broader term/narrower term network) primarily uses a hierarchical structure to represent knowledge. Dewey’s intention was to create a practical tool for librarians for matching the subject content of publications to the classification structure. Thus library classification is more concerned with how concepts are structured in order to group like materials together for easy browsing and retrieval. This means that while library classification takes an integrated approach, it is also practical. A library classification usually is not concerned with whether a concept is covered in a library’s collection, but more
with whether or not the knowledge structure covers all components at each level and all their aspects.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Library classification</th>
<th>XML DTD/Schema</th>
<th>Ontology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purpose</td>
<td>Knowledge structure for organizing library materials</td>
<td>Data model for organizing data</td>
<td>Conceptual model for a knowledge and/or application domain</td>
</tr>
<tr>
<td>Concept specification</td>
<td>Hierarchical</td>
<td>Hierarchical</td>
<td>Hierarchical</td>
</tr>
<tr>
<td>Structure</td>
<td>Class name</td>
<td>Element name</td>
<td>Class name</td>
</tr>
<tr>
<td>Concept labeling</td>
<td>Scope note</td>
<td>Comment</td>
<td>Documentation</td>
</tr>
<tr>
<td>Concept definition</td>
<td>Subdivided-by criteria or facets</td>
<td>Element attributes</td>
<td>Class slots</td>
</tr>
<tr>
<td>Concept attributes</td>
<td>N/A</td>
<td>Text-based</td>
<td>Base data types and complex data types</td>
</tr>
<tr>
<td>Attribute type</td>
<td>See, see also</td>
<td>Entity, ID, IDREF</td>
<td>Inheritance, slot type for class and instance, inclusion of other ontologies</td>
</tr>
<tr>
<td>Relations between concepts</td>
<td>N/A</td>
<td>Character string</td>
<td>SQL compliant, non-SQL data types</td>
</tr>
<tr>
<td>Data modeling</td>
<td>N/A</td>
<td>Relational, Object-Oriented</td>
<td>Relational, Object-Oriented</td>
</tr>
<tr>
<td>Data structure</td>
<td>N/A</td>
<td>Character string</td>
<td>SQL compliant, non-SQL data types</td>
</tr>
<tr>
<td>Data type</td>
<td>N/A</td>
<td>Natural language and/or controlled vocabulary</td>
<td>Natural language and/or controlled vocabulary</td>
</tr>
<tr>
<td>Representation language</td>
<td>Natural language</td>
<td>Natural language and/or controlled vocabulary</td>
<td>Natural language and/or controlled vocabulary</td>
</tr>
<tr>
<td>Definition language</td>
<td>N/A</td>
<td>XML</td>
<td>RDF(S), DAML+OIL</td>
</tr>
<tr>
<td>Markup language</td>
<td>N/A</td>
<td>N/A</td>
<td>First-order logic</td>
</tr>
</tbody>
</table>

Table 1. Comparison of library classification, XML DTD/Schema, and ontology

Note: RDF(S) = Resource Description Framework (Schema); DAML = DARPA R Agent Markup Language; OIL = Ontology Interchange Layer.

Similar to classification, XML schemas organize concepts into a hierarchy, but they are more data-oriented. That is, XML schemas show a very strong tendency for representing concepts involved in an application domain and view these concepts at the logical level. This mandates that an XML schema must provide a conceptual model for a domain by specifying what concepts there are,
what the attributes are for each concept, and in which way concepts are related in an application domain.

The ways ontologies specify concepts are similar to those of XML schemas in that they are both application oriented. Because of this, ontologies in general are not intended for representing the complete human knowledge system but instead, the concepts useful to an application system. A unique method used in ontologies for creating relations between concepts is using complex slot types such as class and instance (Figure 2). Such complex slot types provide deeper representation for the multi-dimensions of concepts. Compared to the other two structures in Table 1, ontologies provide a fuller range of mechanisms for representing and organizing knowledge.

![Figure 2. An ontology example: a class is used as the value domain for another class](image)

**4. Discussion and Conclusions**

Classification, XML schemas, and ontologies share some common approaches to representing and organizing knowledge, but they are produced under different paradigms and serve different purposes. The comparison of these knowledge structures indicates that more recent approaches to knowledge representation and organization are developed using the foundations established by precursors. Classification existed long before the computer was invented. While classifications are still being used and developed, technological advances motivated the evolution of newer knowledge-representation paradigms, which in turn generated new structures, such as XML schemas and ontologies. As shown on the paradigm chart (Figure 1), development in one area may not always move along a single direction as indicated by the direction of the arrows. Any paradigm can also move inward towards the center.

This analysis of KR paradigms is only preliminary. Before we can fully describe the model suggested here, the questions raised need to be studied not only in the context of the examples used in this paper, but also in extended examples from other domains.

**References**

New York: Routledge.
Jens-Erik Mai  
The Information School, University of Washington, USA

Is Classification Theory Possible? Rethinking Classification Research

1. Introduction
Theoretical context independent explanations of classification could enhance the universality of classification research and make knowledge about classification available to settings other than traditional libraries. There is a tremendous need for constructing classificatory structures in a range of settings many of which are far removed from the environment in which classification theory and research has been practiced in the last century and a half. The construction of classificatory structures on the Internet, intranets, and in knowledge management systems has received some attention lately. The question examined here is whether it is possible to create a single theory of classification that applies to the range of contexts in which classificatory structures are applied.

The object of this paper is to question the assumption that bibliographic classification theory can resemble scientific theories. It is argued that the context of any classification influences the use and understanding of the classification to such a degree that the classification cannot be understood separate from its context. Furthermore, the development from being a novice classifier or classificationist to becoming an expert is explored. It is assumed scientific theories must relate as much to the activity of novices as to the activity of experts and that scientific theories explain both what it is that novices do and what experts do. It is argued that expertise is achieved not through a correct application of a classification theory but through experiences and adjustment to a particular context and that the activities of novices are quite distinct from the activities of experts in that experts draws on the context of the situation and that novices do not.

2. Theory of Classification
Langridge (1976) provides an account of the principles of constructing knowledge organization systems and the theoretical underpinnings of different approaches. He identifies four principles that have guided construction of knowledge organization systems: 1) ideological, 2) social purpose, 3) scientific, and 4) the disciplines. The ideological principle organizes knowledge according to an ideology that the knowledge organization system serves. Langridge gives the examples of "the Christian schemes of the Middle Ages and the Soviet scheme which substitutes for the Bible and Christianity the works of Marx and Lenin and the 'religion' of communism" (Langridge, 1976, p. 4-5).

As an example of a knowledge organization system guided by social purpose, Langridge refers to Ranganathan's example of an ancient Indian classification that organizes knowledge "in an order of decreasing immediacy of
social use" (Langridge, 1976, p. 5). The system only has four classes and has not been used to organize bibliographic material.

The last two principles--scientific and disciplines--are the only two principles that have been used in modern classification theory. The latter is Langridge's contribution and the former the most widespread.

Langridge's idea is that is a set of disciplines of knowledge that most philosophers agree on and that these disciplines could be used as the basis of classification. He surveys seven philosophers' (Mortimer Adler, Ernst Cassirer, R.G. Collingwood, Paul Hirst, Charles Hillis Kaiser, Philip H. Phenix, and Louis Arnaud Reid) categorizations of the sciences and finds a similarity between them. Each of these philosophers divides the entire universe of knowledge into 4-7 fundamental disciplines of knowledge. They do not sub-divide the disciplines. Langridge argues that a general classification scheme should use these fundamental disciplines of knowledge to organize the universe of knowledge and then further subdivide the individual disciplines by the phenomena studied or methods used. Langridge notes that it more common to divide according to specialization but that "today it seems likely that it is more helpful to divide by phenomena studied" (Langridge, 1976, p. 9).

Common for these three principles is that they use categorizations that already exist as the basis for developing a bibliographic classification. This basis could either be an ideological categorization, a social categorization, or a philosophical categorization. To make a successful bibliographic classification based on any of these principles it is required that the categorization principle is widely accepted and recognized.

Langridge names the fourth principle the scientific principle. It is probably the principle that has received the most attention in modern classification theory. The principle began with the work by Richardson (1935, first edition 1901) whose view can be summed up by "two statements: the order of the sciences is the order of things, and the order of the things is the order of their complexity" (Langridge, 1976, p. 5). Although the theory applies to natural objects it has not always been made clear, that the theory's relation to man made things was a matter for investigation and should not be taken for granted. Langridge finds that the principle at first sight meets his requirement of neutrality but that "serious objections ... arise from [the] underlying philosophical assumptions" (Langridge, 1976, p. 5). These assumptions are: 1) Science is the paradigm of all knowledge, 2) Knowledge is a simple unity, and 3) There is an objective external world. Langridge dismisses the scientific principle because it is tied to these positivistic assumptions.

3. Positivistic Epistemology

Harris (1986a; 1986b) has investigated the historical roots of the commitment to positivism in LIS. He finds that it can be traced back to the Carnegie Cooperation's establishment of the Graduate School of Library Science at the University of Chicago in 1928. Harris argues that the first faculty at the school defined and established what he at one point calls "a 'psychosociological' research program" (Harris, 1986a, p. 517) and at another point calls "a positivistic epistemology" (Harris, 1986b, p. 218). The school further succeeded in forming a conception in the LIS field over the next several generations that 'good research' is founded in such an approach (Harris, 1986b, p. 218). Harris further argues that the
belief in the possibility and desirability of a science of librarianship recalls the situation of the social sciences in the 1950s (Harris, 1986a, p. 528; 1986b, p. 220-221) summarizes the case against positivism in the social sciences in three points,

1 the social sciences have never been able to generate "scientific paradigms" that might govern research in ways analogous to the hard sciences;
2 the social sciences simply cannot sustain the essential division between the subject and the object of research so central to the positivist epistemology;
3 the value free pretensions of the social sciences have been proven to be mystifications designed to camouflage the extent to which the social scientist is governed by prejudgments and domain assumptions.

LIS must separate itself from a positivistic epistemology in the same way that it has been necessary in the social sciences, even though a positivistic or quasi-positivistic approach remains the most widespread approach to research and thinking in the LIS field. This is seen by the emphasis on objectivity, generalization of findings, empirical research, quantification, etc. Most importantly such an approach to the "field has severely limited the range of questions that can be investigated and has rigidly defined the characteristics of relevant answers" (Harris, 1986b, pp. 221-222).

An ideal theory of classification that follow the neutral, objective and positivistic line of thought should be able to prescribe how a set of documents should be organized and predict the consequences of the organization. The theory should furthermore apply to all kinds of different settings, users, and document types. Flyvbjerg (2001, pp. 38-39) sums up the requirements of an ideal theory based on his readings of Hubert Dreyfus and Pierre Bourdieu. Flyvbjerg finds that ideal theories have six characteristics:

- Explicit. A theory needs to be laid out in such detail that any reasoning human being is able to understand it. The theory must not stand or fall on interpretation or intuition.
- Universal. The theory must apply at all times and in all places.
- Abstract. The theory must not require reference to concrete examples.
- Discrete. The theory must be formulated with context-independent elements; it cannot refer to human interests, traditions, institutions, etc.
- Systematic. The theory must constitute a whole in which the context-independent elements are related by laws or rules.
- Complete and predictive. The theory must be complete in the sense that it covers its whole domain and it must be predictive in the sense that the theory must specify the effects of the elements.

These characteristics cannot, of course, be fully realized in all situations. It is only possible to approach them to varying degrees. However, they do characterize what a theory is and what we can expect from a theory. These characteristics are the frame of reference for understanding the scope and possibilities of classification theory and insofar as a general classification theory should be regarded as a scientific theory about classification of documents, it needs to conform to these characteristics of scientific theories.
4. Context, Theory, and Classification

Everyday practical librarian and information specialist work is based on adaptability to individual work domains, understanding of individual needs, interpretation of documents for representation, etc. all of which takes place in particular contexts. Much research in classification, on the other hand, is based on an assumption that classification research should be objective, neutral, and replicable.

In a personal statement in one of the two 50th anniversary issues of JASIS Michael Buckland talks about two traditions within library and information science, namely the computational tradition and the document tradition. He defines the document tradition, as being concerned "with knowledge, meaning, learning, description, language and ambiguity, therefore, any view of it remains incomplete unless some roots in cultural studies, in the humanities and qualitative social sciences, is acknowledged" (Buckland, 1999, p. 971). It is very difficult to discuss classification of documents without having focus on the meaning of the documents and the language they use. Any investigation into classification must start with a clear and explicit theory of language, meaning, and influence of context on that. Furthermore, any empirical investigation of classification needs to explicitly take into account the context of situation studied. Therefore, classification research must in reality be studies of human activity, language, and interpretation.

Any application of an overreaching bibliographic classification theory will in such practical circumstances fall short. The reason for such a conclusion is that a theory that makes explanation and prediction possible necessarily must exclude the concrete context of everyday activity. However, this exclusion of context makes explanation and prediction impossible.

5. Novice and expert classifiers and classificationists

Classification is primarily a human activity and research and studies of classification is an attempt to understand human activity. One of the ways to understand the limitations of a general classification theory is to see how people learn how to classify. Dreyfus and Dreyfus (1986) have identified five stages of development from being a novice at some task to becoming an expert. People acquiring a certain skill will pass through these stages in becoming better and better at the skill and mastering it in the end. People at one stage will be better at the skill than people at a previous stage. There are virtually no limits to the types of skills that Dreyfus and Dreyfus' stages cover, as long as the skill can be acquire by any rational human being (as for instance bicycling, playing chess, surgery) although not all people will be able to reach the higher stages in all skills.

The five stages and some simple notions of them are:

Stage 1. Novice. "The novice learns to recognize various objective facts and features relevant to the skill and acquires rules for determining actions based upon those facts and features" (Dreyfus and Dreyfus, 1986, p. 21). The novice will only be able to apply the learned rules in contexts and situations with which the novice is familiar. The learned rules are therefore context free rules. The rules are context free in the sense that the novice has learned the rules objectively and will apply the rules to the particular situations with no regard for the situation itself. The novice will apply the learned rules in situations that match the learned proto-situation and
will not be able to accommodate for the given context of a new and different situation.

Stage 2. Advanced beginner. As the novice becomes familiar with the skill, she advances to a marginally acceptable level of performance. This is done "through practical experience in concrete situations with meaningful elements, which neither the instructor nor the learner can define in terms of objectively recognizable context-free features" (Dreyfus and Dreyfus, 1986, p. 22). The advanced beginner begins to recognize and apply previously experienced situations to new situations. Hence the advanced beginner uses both context-free rules and "situational" elements in performing a task.

Stage 3. Competence. Whereas the novice and the advanced beginner merely followed a set of rules in coping with a situation, the competent performer sees a situation as a set of facts. The competent performer approaches the situation with a plan to organize the situation in mind. The situation is then examined according to the plan, and only those factors which are important to the given plan are given consideration. Since the competent performer acts according to a plan which she has chosen herself, she "feels responsible for, and thus emotionally involved in the product of choice" (Dreyfus and Dreyfus, 1986, p. 26). The novice and the advanced beginner, on the other hand, act according to specified elements or rules, and thus feel that an unfortunate outcome is the result of inadequately specified elements or rules.

Stage 4. Proficiency. Unlike the previous performers, the proficient performer does not rely on rules in performing a task. At this stage she merely uses her intuition. This should not be confused with either irrational conformity or guessing. To guess is to reach a conclusion when one does not have sufficient knowledge or experience. Intuition here means the "sort of ability we all use all the time as we go about our everyday tasks" (Dreyfus and Dreyfus, 1986, p. 29). The proficient performer is characterized by her ability almost immediately without conscious effort to follow a plan after encountering a problematic situation. Yet it is unlikely that the proficient performer will be able to explain in detail how she solved the situation.

Stage 5. Expertise. "An expert generally knows what to do based on mature and practiced understanding" (Dreyfus and Dreyfus, 1986, p. 30). The expert does not see problems in some detached way. She is so much a part of the situation that she does not need to be aware of the situation to handle it. "When things are proceeding normally, experts don't solve problems and don't make decisions; they do what normally works" (Dreyfus and Dreyfus, 1986, p. 31). When things work as they normally do, the expert is able to deliberate before acting. But this deliberation is not based on calculative problem solving. Rather, it is based on critically reflecting on her intuitions. Most importantly, as the performer becomes an expert her performance becomes fluid. The expert seldom needs to consider what she does, but when she needs to she is able to do so critically. And, like the proficient performer, the expert will not be able to articulate her exact actions. The expert will likewise not be able to legitimate her decisions, but even so she will seldom make wrong decisions.

Dreyfus and Dreyfus' model contains an important jump between the three first stages and the last two stages. Somewhere around this point in the advancement from novice to expert does the person abandon using rules as the main guidance in performing the task and relies instead on intuition and context. There is,
in other words, a crucial difference in how less experienced people and more experienced people perform a certain skill. Less experienced people more or less apply a set of rules to a given situation whereas more experienced people rely on their tacit experience and intuition.

A theory of classification of documents should inform people at the first earlier stages about how to perform the task as people on the latter stages do. However, people on the latter stages perform the task without being conscious about exactly what it is that they do and they are not able to explicitly formulate what they do. Furthermore, when a person learns how to classify, the person applies some very simple context-independent rules, techniques, and theories about classification but when the person becomes more experienced, he/she will rely less on these rules, techniques, and theories and more on his/her experience, intuition, and the particular context.

6. Discussion and conclusions

The simple conclusion is that it is not possible to develop a general bibliographic scientific classification theory. The reason for this conclusion is that classification is dependent on the particular contexts to such a degree that statements about classification in reality are statements about how human being acts and should act in particular circumstances and context. However, one of the particular requirements to a scientific theory is that it needs to be independent of particular interpretations, contexts, examples, traditions, etc. In other words, the theory must be able to prescribe how the classification should be constructed and predict the consequences of it. Furthermore, classification is a skill that only partly relies on rules, techniques, and theories; expert uses their intuition and experience when classifying. More experienced classifiers performs the task much different from less experienced classifiers. The study of experts will not directly uncover how they perform their tasks such that less experiences classifies could copy the experts' way of doing things.

The upshot of this discussion is a call for rethinking classification research. Classification research is about how human beings act in particular contexts and circumstances and that is what classification research should be about. There does not and cannot exist a general universal theory about classification of documents. The best way to learn about classification is through studies of how and why people act as they do in particular contexts and circumstances. In other words, the case study method seems like an ideal method for classification research. At the center of the case study method lies the study of context-dependent knowledge and expertise.

References
Where Have All the Flowers Gone?
An Investigation into the Fate of Some Special Classification Schemes

Abstract: Prior to the OPAC many institutions devised classifications to suit their special needs. Others expanded or altered general schemes to accommodate specific approaches. A driving force in the creation of these classifications was the Classification Research Group, celebrating its golden jubilee in 2002, whose work created a framework and body of principles that remain valid for the retrieval needs of today. The paper highlights some of these special schemes and highlights the fundamental principles which remain valid.

1. Introduction

The distinction between a general and a special classification scheme is made frequently in the textbooks, but is one that it is sometimes difficult to draw. The Library of Congress classification could be described as the special classification par excellence. Normally, however, a special classification is taken to be one that is restricted to a specific subject, and quite often used in one specific context only, either a library or a bibliographic listing or for a specific purpose such as a search engine and it is in this sense that I propose to examine some of these schemes. Today, there is a widespread preference for searching on words as a supplement to the use of a standard system, usually the Dewey Decimal Classification (DDC). This is enhanced by the ability to search documents full-text in a computerized environment, a situation that did not exist 20 or 30 years ago. Today's situation is a great improvement in many ways, but it does depend upon the words used by the author and the searcher corresponding, and often presupposes the use of English. In libraries, the use of co-operative services and precatalogued records already provided with classification data has also spelt the demise of the special scheme. In many instances, the survival of a special classification depends upon its creator and, with the passage of time, this becomes inevitably more precarious.

2. Disciplines needing special treatment

Certain disciplines have always lent themselves to special treatment for a wide variety of reasons. Those with local peculiarities, such as Politics, Religion or Education are all examples. Bias is a strong element here. Despite the purists, there are occasions when bias is helpful - a Roman Catholic database is not going to favour a Protestant approach - and in religion it is extremely difficult to find a scheme that is bias-free, and it is probably not very helpful if it is. One of the earliest special schemes for Religion was the now largely forgotten Pettee classification (Pettee, 1911). Another example is the Lynn-Petersen scheme (Lynn, 1968) devised to counteract the strong Protestant bias of DDC and LCC, at least as
they appeared in the 1930s, which is now rarely used, apart from in such libraries as the Jesuit foundation Heythrop College in the University of London where it is still applied. Similar special schemes, such as those of Elazar (Elazar, 1968) for Hebraica or Ziauddin (Ziauddin, 1979) for Islam take care of those religions which receive less satisfactory attention in the widely-used general schemes.

Before the days of cataloguing co-operatives, local expansions of general schemes were a popular phenomenon. Classification schemes were expanded in-house to accommodate local requirements. Area Studies is one example. Where an approach by area, subdivided by topic is needed, a simple solution is to use a set of geographical subdivisions preceding the subject subdivisions of a general scheme. The School of Oriental and African Studies in the University of London is just one example of a library which is arranged in this way taking its own set of areas, followed by DDC subdivisions. Another popular solution is to produce an expanded schedule for the topic of interest. The Oxford Forestry Classification (IUFRO, 1990) is an instance. This classification, is an expansion of UDC's class 63, originally for use in the University of Oxford. It was then published as an authorized expansion of the UDC the erstwhile full schedules and gained widespread acceptance. Although no longer a recognized part of the UDC, and maintained in Vienna, it continues to be used across Europe.

Law is another subject where special schemes have been preferred. The reason for Law being different, apart from an ingrained preference by lawyers to prefer broad subject arrangements based on an alphabetical array of topics, is the lack of a general scheme to meet their needs until comparatively recently. In the UK and through much of the Commonwealth, where the legal system is comparable, the choice has often been the Moys classification, which last year achieved its fourth edition. But it is questionable, now that its originator has died, how long the scheme will remain viable - perhaps a decade, unless some organization has the funding and expertise to maintain it in competition with the Library of Congress Class K schedules whose classmarks now appear on all cataloguing data.

Moys' scheme was created originally for use in Nigeria, at Lagos University Library, classified by LCC, which at that time could not accommodate British or Commonwealth (or indeed any other) law. It draws upon much of the fundamental theory that was being refined in the 1950s and 60s, especially by the Classification Research Group (CRG), whose contribution will be examined later. Law is divided by legal system then legal subject (the reverse of the then DDC order) and a series of auxiliary tables is provided to accommodate many of the recurring facets. In later editions it also draws upon the work of one of the CRG's members, Jean Aitchison, by adding a thesaurus type index. A further benefit is that it provides two different types of notation, one of decimal numbers using 340, to take the place of DDC's 340 class, and one of letters and numbers, beginning with the letter K for the core subject, to fill in the gap of the then non-existent Class K in the LCC.

3. Role of the Classification Research Group (CRG)

2002 sees the golden jubilee of the Classification Research Group. Fifty years ago this group of people who were to have a great influence upon research into knowledge organization, as it is known today, first began to meet regularly and to experiment with the creation of classifications to satisfy the needs of their
particular individual working environments. It has become customary to refer to the Group as if it was, and indeed still is, a coherent whole sharing exactly the same views and opinions. Nothing could be further from the truth. The Group was assembled to devise systems for the post-war organization of information retrieval in the UK and its membership was drawn from a range of different library backgrounds. The needs of those diverse backgrounds were capitalized upon and over the course of some 20 years its members honed their conclusions into a coherent body of theory that remains both valid and important for the retrieval needs of the 21st century. Many of these people figure sparsely in the citation indexes that are taken so seriously today. They were far more influential through their work as teachers, as editors of the national bibliographies which were appearing in a flourish of post-war bibliographical activity and as compilers of special schemes to cope with the detailed materials they were having to handle in their daily work.

National bibliographies provide one clear instance of the environment in which the members worked out many of their ideas. The British National Bibliography (BNB) was founded in 1950, using the 14th edition (1942) of the Dewey Decimal Classification, as the basis of its arrangement. Its original editor was Jack Wells, a member of the CRG who was joint author of one of the first textbooks on library classification that directly reflected Ranganathan's teachings, The fundamentals of library classification (Palmer and Wells, 1951). Almost a decade later the BNB transferred its allegiance to the 16th edition of DDC, but it still found that scheme inadequate both in its American approach, and in its lack of detail for modern advances in technology, to act as an efficient retrieval tool. The classified sequence was enhanced by a chain index (invented by Ranganathan), an indexing technique widely used to complement the classified catalogues which were then standard in British public libraries.

It was decided to devise a set of special schedules (BNB, 1963), using a lower case alphabetical notation, to accommodate those concepts where DDC was felt to be wanting, or to take care of such disciplines as Politics or Education, where DDC had an unhelpfully American focus, and to provide the expanded detail needed in Technology. These expansions, which remained in use for over a decade in the 1960s and early 70s, provided a fully faceted approach. They also included a revised auxiliary table of common forms, more logically arranged than in the parent scheme, and with the additional enhancement of making provision for phase relations. This need was not then specifically recognized by any general scheme, the UDC's colon being "all things to all men" and LCC's "general special" subdivisions being far from adequate for this purpose.

Some seven years after the BNB's inception, it was decided that there should be a separate listing for music, the British Catalogue of Music, and here the DDC was found even more wanting. The majority of British libraries favoured a more detailed classification, based broadly on Dewey, known as the McColvin classification after its inventor, the City Librarian of Westminster, which in its Central Music Library housed one of the largest music collections in a British public library (McColvin, 1965). So a member of the BNB's staff, who was also a member of the CRG, Eric Coates, was given the task of devising a totally new classification (Coates, 1960). This is a fully faceted scheme, with an alphabetical notation using capital letters which did not reflect hierarchy. This conformed to the then received doctrine that the function of notation was simply to mechanize the
order, and that such extras as reflecting structure were luxuries that could not be afforded and which would not endure with the advance of knowledge and the need to insert new concepts. This is a point of view which has been discredited in more recent years, where the virtues of an hierarchical notation, as recognized by Dewey, to represent the classification’s arrangement, have become evident in the use of searching via notation in a networked environment.

It was not solely in the realm of national bibliographies that the CRG exerted its influence. Many of its members were working in organizations such as the Metal Box Company, Tate and Lyle, the Aeronautical Research Establishment, the English Electric Company and similar specialized institutions, where existing general schemes failed totally to supply detail of an adequacy to provide a useful retrieval tool for the indexing of highly specialized technical materials. Even where such general schemes did have this facility, for instance the then full tables of the UDC, the emphasis was frequently not where the institution wanted it. This situation still obtains today when the specialist is seeking material, whatever source, printed or digitized, is being used. The 1950s and 1960s consequently saw a proliferation of special schemes produced in the UK for one particular institution or bibliographical listing, and designed totally with the specific needs of that organization in mind. It also saw the testing of many theories that today are taken for granted when framing tools for retrieval, but which at that time had not been widely disseminated beyond the Indian subcontinent.

Many of the schemes on which the Group expended its energies were for technical libraries, but other disciplines were not ignored. The Social Sciences were thoroughly examined by Barbara Kyle, and she was responsible for the schemes used in the 4 Unesco Social Science Bibliographies which covered Politics, Economics, Sociology and Social Anthropology. Experience with these indexing services led her on to draw up the Kyle Classification for the Social Sciences (Kyle, 1960) which was never completed before her death in 1966. This experiment also led to the use of “levels of integration” as the basis of a classification, which were later featured in other enterprises undertaken by the Group.

One specific Social Science where neither DDC nor LCC was seen to be appropriate for British requirements was Education. Consequently, many of the post-war Institutes of Education opted for the Bliss Bibliographic Classification, which they found more adaptable for the British educational system. Not so that at London University, however, where Foskett's *London Education Classification* (Foskett, 1974) still holds sway, though its pronounceable notation, another CRG experiment in notation, had to be abandoned in its 2nd edition. This edition was enhanced by a thesaurus form index. The CRG also produced a classification for Library Science in the 1960s which is of interest because the citation order was changed between the first and second editions. The result was that the same scheme, but with a different facet order, was used in the Library Association's library and in their indexing journal, *Library and Information Science Abstracts* for a number of years.

After some 10 years of experimenting with special schemes, the CRG obtained a grant in 1963 to develop a general classification, under the auspices of NATO - a prelude to the more ambitious UNISIST programme of UNESCO in the 1970s which saw the production of yet another scheme, the *Broad System of Ordering* (Coates, 1978). The NATO grant gave the members of the CRG an opportunity to refine their experiences with creating special schemes into a standard
set of principles appropriate for the construction of a new classification to comprehend the whole of knowledge. This they never achieved, but there are two legacies of the work that was undertaken at that time. The first is the PRECIS system of indexing (Austin, 1974), now fallen into disuse, but much promoted in the 1970s and 1980s. This was devised by the second of the CRG’s two research assistants, Derek Austin, and was used for indexing the BNB for over 15 years, as well as by other institutions, such as the Canadian Film Institute. The high cost of its application rather than any unreliability of the system were its downfall - it is expensive and time-consuming to apply and requires well trained, competent indexers. The lean years of the later 1980s and the widespread introduction of computers at the same time, together with the attractions of easy searching on words in titles caused its demise. This was a disaster, since PRECIS is highly appropriate for use in computerized systems and provides the facility for clear and precise specification.

The other legacy of the CRG’s work is the 2nd edition of the Bliss Bibliographic Classification, on which much of its energies over the past twenty years has been spent. In terms of the theme of this paper, the Bliss Classification is an anomaly. It has many claims to be considered the offspring of the special schemes of the 1950s and 1960s, not altogether surprisingly since it is master-minded by Jack Mills who has dominated much of the Group’s thinking over the past half century. But it is a general scheme, or when completed will be one, though it is not widely used apart from enjoying a measure of popularity in the libraries of the University of Cambridge. This second edition of the classification which bears little resemblance to the first except in very broad outline, embodies many of the special schemes that were devised in those palmy days. Its table of standard subdivisions, which is by far the best in any general classification scheme, is directly derived from the old BNB schedules. Standard subdivisions are a major problem for editors of classification schemes. They tend to become set in stone, so once an arrangement has been put in place it is very difficult to alter anything, because users will have applied them right across their collections, resulting in widespread alterations which are unpopular for obvious reasons - you can now change records globally, but we have yet to devise a means of doing so on the spines of books! Consequently, many of the general schemes used today for a far greater range of purposes than the original intention of shelf arrangement, have common form divisions which originated in the 19th century and have had to be bent and twisted to accommodate the needs of the subsequent 140 years. Not so Bliss, which provides a better arrangement than that found elsewhere for commonly recurring concepts and has the facility to identify the different types of relation that occur in literature, in a manner not dissimilar to the provisions of PRECIS.

4. Lessons for today and application to the thesaurus

In many ways, the situation 50 years ago (or even 150 years ago) was not very different from that of today. The people working in the information field at that time were faced with the task of organizing great quantities of material, frequently dealing with a high level of technical detail, and had at their disposal classification schemes that were totally inadequate, so they had to improvise. It all seemed very daunting and there were no computers, or only very rudimentary ones. Now, many similar isolated attempts to produce methods of retrieval that are
appropriate to finding the material relevant to specific needs from the mass of varied data that we call the Internet are made. The difference between then and now is that today there tends to be little co-ordination of effort with a resultant diffuseness so that, if meetings such as this do not address it, a unified solution will evade us and we shall not achieve our desired goal. The CRG, in its heyday, met regularly and shared ideas and solutions to very individual problems. Its minutes enjoyed a wide circulation and generated considerable correspondence from overseas members. The Group eventually achieved a unity of structure and an agreed technique and body of theory, despite the very disparate views of its individual members. These theories were disseminated through teaching, for many of the members taught in library schools, and consequently their ideas reached a far wider range of practitioners than simply those employed in bibliographic services departments. They also wrote textbooks and many of their views are now set out in other people's textbooks as irrefutable fundamental theories on which all classificationists should proceed - the introduction to DDC is just one example of how accepted their views have become, in its recommendation of the "standard citation order" in its "Tables of precedence".

Today, we rely for the organization of our libraries and databases on standardization. This saves time and effort and above all money. Records are created once and for all, instead of hundreds of times all round the world. This clearly has great economic advantages. What is lost is attention to the needs of the specialist, and the individual tailoring that is abandoned in favour of off the peg solutions - cost efficient and excellent, in the short term. But the specialist does have special needs and much of the effort of 40/50 years ago that was put into attempting to meet these has been drowned in the overwhelming tide of information and quick-fix solutions. These special needs have always been evident in certain disciplines, as has been noted. Outstandingly, Medicine is the one that has bucked the trend. Forty years ago the majority of medical libraries in the UK used the UDC, some used Barnard's classification, but now, almost without exception, they all use the National Library of Medicine Classification - for precisely the same reason as general libraries use DDC or LCC. A similar example is the SfB (SfB, 1973) for Building, whose continued survival in architects' offices owes much to the fact that professional literature is issued preclassified by the scheme.

What has happened to the other special schemes on which so much attention, argument and discussion were spent? Have they all been totally forgotten. The answer is not entirely. Some are still in use. A sizeable number of law libraries are applying Moys classification. It has been seen that the NLM now dominates the medical scene, though perhaps for reasons other than its intrinsic theoretical perfection as a classification scheme. Coates's music classification can be discerned in the revised 780 schedules of the DDC. A comparison of the structure, not the notation, will reveal that the two are in essence one. The BNB expanded schedules have been extensively used in the 2nd edition of the Bliss Bibliographic classification, as have some of the other schemes produced by members of the CRG such as Foskett's London Education Classification. Another example of survival is Jean Aitchison's Classification for the English Electric Company. This went into four greatly expanded editions, the final one in the format which she christened Thesaurofacet (Aitchison, 1969), a format imitated by Moys in her Law scheme, and by Foskett for the 2nd edition of his Education Classification.
The Thesaurofacet format is that of a faceted classification scheme accompanied by a thesaurus, which acts as the index to the classification or can be used independently as a thesaurus. The thesaurus format has become the most popular retrieval tool of the present day. A means whereby the user can search using his or her own words, rather than having to rely upon some structure based on theories which are not self-evident to the uninitiated and which bear no resemblance to the individual's personal approach, is seen as the best solution to today's needs in a networked world. It is especially recommended if it permits searching on a mixture of the controlled vocabulary of the thesaurus and free text based on the words of a title and abstract or even the complete text of a document. A controlled vocabulary, of which the thesaurus is one example, needs to be based on a proper sound structure, and Aitchison et al.'s work on Thesaurus construction (Aitchison, 2000) has become the standard manual for those who wish to develop this form of retrieval tool.

The special classifications which were so prevalent fifty years ago have been replaced by search engines, ontologies and taxonomies devised to assist in retrieving information from the Web. The work of many librarians went towards the formation of a sound body of theory which has much to offer to those who today are seeking to make some sort of order from the chaos. The achievements of these people, and especially of those who made up the membership of the CRG in its early days, should not be forgotten. Many of the theories that were refined then remain valid, and the specialist schemes can still offer assistance when attempting to create tools to retrieve information on individual disciplines in the 21st century.

References
Abstract: Studies on ethical aspects of our profession rarely focus on matters related with the organization and representation of knowledge, but are directed instead toward such subjects as intellectual property, right to privacy, intellectual freedom, or proper professional conduct. Nonetheless, the technological possibilities nowadays have meant a radical change. In the past, a certain policy for indexing or a classification system produced effects only in the relatively limited setting of a library or information center; but now the indexing or classification of certain electronic information resources has effects that go far beyond the physical boundaries of such institutions, or even those of a country. The objective of the present study is, on the one hand, to identify the principal ethical values related with the organization and representation of knowledge, and on the other hand, to see to what degree they are addressed by the ethical codes of professional associations.

1. Introduction

In today's increasingly complex and multicultural information society, problems related with the access to and use of information take on a critical importance that is not always acknowledged and valued. To know how to create, organize, represent, find, access and use information effectively equips information professionals with tremendous power, and at the same time carries a huge legal and ethical responsibility. Although ethical matters that affect the actions of information science professionals have always been present and taken into account, the development of information and communication technologies, together with the growing process of economic globalization, have modified the context in which these professionals act. Therefore, a reassessment of the situation is needed.

Traditionally, studies on ethical aspects of our field have been centered on professional practices (malpractice, liability...) or else on problems related with different rights: the right to privacy, copyright, intellectual freedom and censorship, the right to access to public information, etc. Very little attention was paid to the ethical matters touching on knowledge organization and representation, for example those deriving from indexing and abstracting policies. A good example of this lack of attention can be found in the thorough and excellent work of Thomas Froehlich (1997) who, nonetheless, dedicates just two pages to problems surrounding the organization and representation of knowledge, in particular with regards to the "classification, cataloging, abstracting and indexing policies and implementations."

But the current technological possibilities have turned the situation around: that set
policy governing indexing or classification practices whose effects were once limited to the confines of the library, now produces false drops, missed documents, or bias of a political, cultural or religious nature that can easily go beyond the physical boundaries of the information center or even the country. As pointed out by Orick (2000), the evolution toward digital libraries is changing the location of the control room, as the collection of the library flows outwards from the library building itself.

The main objective of the present study is to identify the framework of ethical values on which the information professional should base all activities involving knowledge organization and representation. Once identified, we shall try to determine to what extent the deontological codes of our profession expound these values, and what gaps and deficiencies exist. Although it is evident that ethical codes are neither the panacea nor the only tool available to regulate the development of a profession, they constitute an excellent frame of reference for guiding professional behaviour (Oppenheim and Pollecott, 2000). Moreover, as Capurro suggests (2000), in an interconnected world, the place of morality seems—paradoxically—to be more basic than legal norms.

2. Methodology

The organization and representation of knowledge are implicit in most every undertaking in our professional field. They have a transversal nature, which makes it very complicated to isolate those ethical concerns that are most specifically thereto related. In fact, most studies on this subject either revolve around the identification and description of the areas such as intellectual freedom and censorship, intellectual property rights, privacy, right of access to information, etc., or else focus on the analysis of one of them specifically. Precisely for this reason, we take as a departing point a number of recent studies—Rubin and Froehlich (1996), Fernández-Molina (2000), Guimarães (2000), and Koehler & Pemberton (2000)—that center on the different areas and elements of an ethical nature pertaining to our professional area.

As an initial result of analysis, we have identified seven values related with the tasks involving knowledge organization and representation, some of a more general character, others more expressly related.

A. **User interests come first.** This is a very general value, meaning that the activities in knowledge organization and representation should be designed and carried out with the interests of the user in mind, above considerations of the information professionals.

B. **Provide services objectively, with no bias of any sort.** It is necessary to guarantee the objectivity of the vocabulary and other means of knowledge representation used, and of the models and structures adopted for its organization.

C. **Provide users with the most up-to-date and precise information possible.** We could reflect here on a new dimension of Cutter's principle of use, according to which indexing should contemplate the variants in user vocabularies, given that indexes aim to establish a bridge between the search vocabulary of the user and the document. Too often do we come across that very unethical practice of resorting to terms that do not
adequately represent the contents of the document but do make a web page more frequently visited, promising more income.

D. Avoid censorship in the selection of information materials. Although this problem has traditionally been tied to the selection of materials, not to their organization and representation, the new systems of classification and rating of Internet contents have modified that perception. The systems and the filtering software used by libraries and other information centers are actually determinant of the information materials that the user may or may not access. It is obvious that the establishment of the labelling vocabulary, the criteria for assigning labels and the rules for filtering the information are not neutral tasks, but rather imply making judgements of a moral nature.

E. If any type of censorship or filter exists, inform the users of their limitations. This value complements the above one, as the user can be seriously misled by not knowing that he/she cannot access one specific type of information due to the filtering systems.

F. Separate personal beliefs from professional services. This distinction is vital for carrying out the tasks of objectively representing knowledge, and establishing a specific connection between the structures of knowledge of the document or the information and those of the user.

G. Maintain professional competence. Also in this case we are dealing with a very general category, which of course is fully applicable to the activities of knowledge organization and representation.

Having identified the ethical values, we proceeded to choose the codes of conduct of professional associations to be analyzed. In an attempt to have the most representative sample possible, we account for a variety of economic, geographical (Europe, North America, Latin America, Asia, Oceania), cultural and legal areas (Anglo-Saxon and Latin traditions), as well as the different perspectives of the professional sectors represented. For this reason, associations of librarians, special librarians, records managers, archivists, information scientists, and information managers were included in the study. A total of 16 professional associations were finally chosen, and the Web pages containing their ethical codes were consulted. The name of each association, the denomination of its ethical code, the day of its last update and the Web address where it can be located appear in the appendix at the end of this paper. There is one case deserving special comment here: the Canadian Library Association has not only an ethical code but also a series of position statements that are complementary. For this reason, we did not simply refer to their ethical code per se (very brief and not updated since 1976), but also considered the position statements under the headings “Intellectual freedom” (1985), “Information and telecommunication access principles” (1994) and “Internet access” (1997).

3. Results and discussion

The result of the analysis can be seen in Table 1, where the seven values studied appear on the one side, and on the other side we give the codes of the 16 associations (identified by their acronyms). It is evident at first sight that those values of a general nature are addressed by all or nearly all of the codes, whereas more specific values are hardly present. Thus, it does not come as a surprise that the only value included in all the codes refers to the need to maintain professional competence, a basic principle of any profession. Also addressed by the wide
majority is the initial value identified: “the user comes first”, considered by Dole and Hurych (2001) to be the outstanding value of our profession.

<table>
<thead>
<tr>
<th>ETHICAL VALUES</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>1.1.1.1.1</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>CODES OF ETHICS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALIA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>APBAD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARMA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASIST</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBPQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JLA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIANZA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Ethical values and their inclusion in the codes of ethics

In most cases, the value of separating personal beliefs from professional practices is made explicit. Very noteworthy, however, is its absence from the codes of associations with great weight and tradition in the field, such as the Library Association or the Canadian Library Association. Another intrinsic value of our profession, intellectual freedom and the struggle against censorship, appears in most codes, though again its absence in the LA and the New Zealand codes is surprising; and it is difficult to understand why it escapes mention in a code as complete and detailed as that of the Corporation des Bibliothécaires Professionnels du Québec (CBPQ). Meanwhile, the importance of providing an objective, unbiased service only appears in print in about the half of the codes analyzed.

Special mention should be given to two other values, both very closely related with the rise and development of information and communication technologies -and probably for that reason, scarcely present in the codes examined. The first, that information should be as up-to-date and precise as possible, is only specified in the codes of the Portuguese Association (APBAD), the MLA, the CBPQ, and that of information managers (ARMA), which comes as quite a surprise if we bear in mind that it is a consubstantial value within a profession dedicated to the diffusion of information. Even more scarce is the final value considered, that related with censorship and the filters used for this purpose. Only the Portuguese Association and the CBPQ include mention of it. In fact, the CBPQ code establishes in its article 10 "si les téléressources sont filtrées dans le milieu où il oevre, le bibliothécaire doit prendre des dispositions pour que la clientèle soit informée de la
nature et des motifs du filtrage pratiqué.” The ethical codes of these two associations, APBAD and CBPQ, are clearly the most complete ones. The Portuguese code is quite recent and has very complete and balanced contents, the fruit of a long and complex study involving professionals, students of the field, and users of information services. Indeed, it is the only one containing all seven of the values analyzed. On the other hand, that of the library science professionals of Quebec is not only up-to-date but also the longest and most detailed (48 articles), though one wonders why it makes no clear reference to intellectual freedom.

4. Conclusions

The results of our study evidence that, in general terms, the deontological codes of our profession address the principle ethical values related with knowledge organization and representation, though with little specificity, due to the traditional separation between the activities of processing and management. However, there is one blatant deficiency rooted in the advance of new technologies: the problems surrounding the systems of classification and the filtering of Internet contents, especially owing to the confusion produced between selection and information organization and representation. In the recent past, the information professional was the only responsible for the selection and organization of library resources. Now, however, the filters mean that other parties (unknown, and often lacking adequate training) carry out these tasks in a more or less arbitrary manner, imposing their own criteria on the staff of the library or information center, and therefore on the community served as well. Although we might imagine these problems could be solved by establishing rules that regulate intellectual freedom and are opposed to censorship (included in practically all the codes), it is best to make them explicit, as the CBPQ and APBAD effectively do.

The rest of the non-ethical activities carried out in the context of technological development can be faced with the help of our traditional rules and guidelines, as they may be new versions of old problems with which information professionals have always had to struggle.

References


**Appendix: List of professional associations and their codes of ethics**

- Colegio de Bibliotecarios de Chile (CBC). *Código de ética* (1977), http://www.bibliotecarios.cl/etica.htm
- Colegio de Bibliotecarios de México (CBM). *Código de ética profesional*
- Canadian Library Association (CLA). *Position statements*, http://www.cla.ca/about/poslist.htm
Ali Asghar Shiri, Crawford Revie and Gobinda Chowdhury
Department of Computer and Information Sciences, University of Strathclyde, Glasgow, UK

Assessing the Impact of User Interaction with Thesaural Knowledge Structures: a Quantitative Analysis Framework

Abstract: Thesauri have been important information and knowledge organisation tools for more than three decades. The recent emergence and phenomenal growth of the World Wide Web has created new opportunities to introduce thesauri as information search and retrieval aids to end user communities. While the number of web-based and hypertextual thesauri continues to grow, few investigations have yet been carried out to evaluate how end-users, for whom all these efforts are ostensibly made, interact with and make use of thesauri for query building and expansion. The present paper reports a pilot study carried out to determine the extent to which a thesaurus-enhanced search interface to a web-based database aided end-users in their selection of search terms. The study also investigated the ways in which users interacted with the thesaurus structure, terms, and interface. Thesaurus-based searching and browsing behaviours adopted by users while interacting with the thesaurus-enhanced search interface were also examined.

1. Introduction

The last decade has witnessed the emergence of a broad range of applications for knowledge structures in general and thesauri in particular. A number of researchers have predicted that thesauri will increasingly be used in retrieval rather than for indexing (Milstead, 1998; Aitchison et al., 1997) and that their application in information retrieval systems will become more diverse due to the growth of full-text databases accessed over the Internet (Williamson, 2000). Some researchers have emphasised the need for tailoring the structure and content of thesauri as tools for end-user searching (Bates, 1986; Strong and Drott, 1986; Anderson and Rowley, 1991; Lopez-Huertas, 1997) while others have suggested thesaurus-enhanced user interfaces to support query formulation and expansion (Pollitt et al., 1994; Jones et al., 1995; Beaulieu, 1997). The recent phenomenal growth of the World Wide Web has created new opportunities to introduce thesauri as information search and retrieval aids to end user communities. While the number of web-based and hypertextual thesauri continues to grow, few investigations have been carried out to evaluate the ways in which end-users interact with and make use of online thesauri for query building and expansion. The work reported here expands on a pilot study (Shiri and Revie, 2001) carried out to investigate user – thesaurus interaction in the domains of biology and veterinary medicine.
2. Objectives
The main objective of the present study is to evaluate end-users' interactions with an online thesaurus for selecting search terms for query formulation and expansion, specifically:
- to study the patterns of user behaviour in thesaurus-based browsing and searching;
- to provide a framework for assessing the impact of thesaurus interaction on term selection;
- to explore ways of enhancing user interfaces through the use of knowledge structures embedded within thesauri.

3. Methodology

The Information retrieval system. The web-based interface to the CAB Abstracts database provided by Ovid Technologies was used in this study as it provided the system features required for this experiment. The system provides a thesaurus-based searching facility based on the CAB thesaurus in its advanced search mode which maps users' search terms to thesaurus descriptors. It also caters for browsing and selecting terms during the query construction process.

The subjects. The purpose of the study requires the participation of genuine users with real information needs. Faculty and researchers were selected from the departments of veterinary medicine and biology at Glasgow University as these two subject areas are well covered in the CAB Abstracts database. In order to inform and validate the proposed methodology a small pilot study involving four subjects was carried out.

Search requests. Search requests were elicited based on information needs of the researchers prior to running the experiment. Each researcher was asked to provide three search topics of interest. This decision was made on the assumption that evaluation of search term selection can effectively be carried out only if users having genuine information requests take part in the study.

Data gathering techniques. Due to the complex nature of capturing data on all aspects of user-system interaction, and in particular search term selection and thesaurus interaction, this study employed a combination of data collection techniques to effectively capture qualitative as well as quantitative data. A pre-search questionnaire, screen-capturing software, the 'think-aloud' technique, and post-search interviews were all used to collect data at various stages of the users' interaction with the system.

4. Results

To identify thesaurus-based search behaviours and patterns, a number of process measures were used to quantitatively analyse all individual search process characteristics. These measures are as follows: state, move (step), and search term.

State: major stages or conditions a user goes through while conducting a search;

Move (or step): characterises any action a user takes while interacting with the system;
Search term: a general characterisation of all types of terms used during the entire search, provided either by the user or the system.

The results obtained from the small pilot study are summarised in terms of these key process measures.

4.1 Search states and moves

Five main states were defined as characterising a typical end-user search interaction through the thesaurus-enabled interface. Specific moves were identified to describe the steps and actions taken by users within these main states and these are listed in Table 1. The 12 searches conducted by the four researchers were analysed using this framework of states and moves to shed light on various aspects of browsing and searching behaviours, and the results related to moves within the search process are provided in Table 1.

<table>
<thead>
<tr>
<th>Moves</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
<th>S7</th>
<th>S8</th>
<th>S9</th>
<th>S10</th>
<th>S11</th>
<th>S12</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term input (C)</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>35</td>
</tr>
<tr>
<td>Perform search (S)</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>42</td>
</tr>
<tr>
<td>Browse initial mapped terms (C)</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>7</td>
<td>5</td>
<td>38</td>
</tr>
<tr>
<td>Browse thesaurus (C)</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>26</td>
</tr>
<tr>
<td>Select terms (C)</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>5</td>
<td>42</td>
</tr>
<tr>
<td>Combine terms (S)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>22</td>
</tr>
<tr>
<td>Continue search (S)</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>3</td>
<td>8</td>
<td>4</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>64</td>
</tr>
<tr>
<td>Search reformulation (C)</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>Browse titles (C)</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>View complete records (C)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>Total moves per search</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>317</td>
</tr>
</tbody>
</table>

Table 1. Moves of different types summarised for all searches

Moves were assigned to one of two main categories to aid the analysis of user interaction behaviour in more detail:

a) "conceptual moves" (C) in which users perform some kind of conceptual analysis of terms or documents;

b) moves associated with using system features (S) such as performing a search or;

c) combining terms.

Users made a total of 317 moves of which 189 (60%) were of a conceptual nature, the rest being associated with the use of system features. On average users made 26.4 moves per search. Conceptual moves can be broken down into three
types i.e. moves related to entering search terms, moves dealing with thesaurus browsing and term selection or query reformulation, and moves connected with the retrieved results such as browsing titles and viewing complete records. Of the 189 conceptual moves 106 (56%) were associated with browsing the thesaurus, viewing mapped terms, and term selection. It is also evident that around one third of all moves were connected with browsing and choosing search terms from mapped and thesaurus terms.

4.2 Search term selection

In total users browsed 710 descriptors, selecting 82 of these for use during the 12 searches. Thus on average around 12% of descriptors browsed were actually selected though there was clearly significant variability, with specific term selection figures ranging from 5% to 29% amongst searches. Summarising the measures on a search-level basis indicated that on average around 3 terms were initially provided; 60 terms viewed; and ultimately around 7 terms selected for inclusion in the search.

This 'average behaviour' could suggest a three parameter thesaurus interaction model of the type:

\[ f(\text{initial terms entered, terms viewed, terms ultimately selected}) \]

where, assuming \( N \) is the number of initial terms, then the number of terms viewed and the number of terms selected would be roughly equal to \( 20*N \) and \( 2*N \) respectively. However, these average figures obscure a wide variety of interaction behaviours with users viewing between 6 and 132 terms over the 12 different searches and ultimately selecting anything between 1 and 22 terms. A better approach might be to compare search characteristics according to their membership of some categorical groupings, for example: few initial terms and a low level of interaction, or many initial terms and a high degree of interaction.

An attempt to formulate a more flexible model is shown in the Thesaurus Interaction Impact Matrix (TII2M) in Figure 1, which categorises searches based on the number of initial search terms used and the impact that thesaurus interaction had on the number of the terms selected. (The average number of terms viewed for each quadrant of the model is also shown).

It can be seen that those searches where a user had a larger number of terms in mind as they came to carry out a search, resulted in on average three times as many terms being viewed. However, in the case of both the "many" and "few" initial terms categories the relative number of terms eventually selected was not related to the number of terms viewed. It is anticipated that the use of the Thesaurus Interaction Impact Matrix (TII2M) may aid the modelling of interaction, browse and navigation behaviours adopted by different users in their formulation and expansion of queries.

4.3 Users' general impressions of the thesaurus and its interface

The qualitative judgements of users were also collected to evaluate the usefulness of both the thesaurus and the interface. All users stated that the thesaurus provided them with additional search terms to choose from. It was seen to be useful when the user entered a term without any confidence and the mapped term was the exact term the user was looking for. It was commented that the thesaurus would certainly provide additional search terms and options for narrowing down the search and it gives a different perspective of the same subject. The users were also asked if there were additional search terms provided by the thesaurus, which they were not aware of at the beginning of search. All users stated that there were new search
terms that they would not think of at the outset. They also pointed out that the thesaurus was helpful in dealing with variant spelling and word forms.

<table>
<thead>
<tr>
<th>Effect of thesaurus interaction on term selection</th>
<th>Initial term entry</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Term selection</td>
<td>Few terms (2 or less)</td>
<td>Many terms (more than 2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of terms remains similar</td>
<td>Average number of terms viewed = 29</td>
<td>Average number of terms viewed = 93</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S1</td>
<td>S7</td>
<td>S1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S2</td>
<td>27</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S6</td>
<td>53</td>
<td>97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>6</td>
<td>109</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S4</td>
<td>5</td>
<td>S8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S5</td>
<td>7</td>
<td>S10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S6</td>
<td>27</td>
<td>80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S7</td>
<td>97</td>
<td>132</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S8</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S9</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S10</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S11</td>
<td>132</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Search number</td>
<td>Terms viewed</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Users were also asked to comment on the user interface to the thesaurus and its search and browse functions. Check boxes and hypertext features associated with thesaurus descriptors together with the combine search feature were mentioned by all users as easy to use and intuitive. Simple issues such as the location of the continue search command, one of the most frequently used system feature, was a problem as users search throughout the whole page to find this icon.
5. Conclusion
This paper has reported a quantitative analysis framework for the evaluation of users' interaction with an online thesaurus and the ways in which it affects their search term selection behaviour. The framework suggests a basis for detailed analysis of steps and moves users adopt during a thesaurus-based search session within the context of a web-based information retrieval system. A Thesaurus Interaction Impact Matrix (TI²M) has been developed to provide an approach to modelling the interaction of different user groups with the thesaurus using quantitative analysis of the number of search terms entered, viewed and selected.

A larger study is currently being undertaken to evaluate detailed user-thesaurus interaction and to find possible thesaurus browsing and searching behaviours, as well as issues affecting users' attitudes in selecting search terms for query formulation and expansion.

Acknowledgements
The authors acknowledge CAB International for providing access to the Ovid CAB Abstracts database on the web as the experimental platform used within this study.

References
Ariadne's Thread: Knowledge Structures for Browsing in OPAC's

Abstract: Subject searching is the most common but also the most conflictive searching for end user. The aim of this paper is to check how users expressions match subject headings and to prove if knowledge structure used in online catalogs enhances searching effectiveness. A bibliographic revision about difficulties in subject access and proposed methods to improve it is also presented. For the empirical analysis, transaction logs from two university libraries, online catalogs (CISNE and FAMA) were collected. Results show that more than a quarter of user queries are effective due to an alphabetical subject index approach and browsing through hypertextual links.

1. Introduction

Since the 1980's, online public access catalogs (OPAC's) have become usual way to access bibliographic information. During the last two decades the technological development has helped to extend their use, making feasible the access for a whole of users that is getting more and more extensive and heterogeneous, and also to incorporate information resources in electronic formats and to interconnect systems. However, technology seems to have developed faster than our knowledge about the tasks where it has been applied and than the evolution of our capacities for adapting to it. The conceptual model of OPAC has been hardly modified recently, and for interacting with them, users still need to combine the same skills and basic knowledge than at the beginning of its introduction (Borgman, 1986, 2000): a) conceptual knowledge to translate the information need into an appropriate query because of a well-designed mental model of the system, b) semantic and syntactic knowledge to be able to implement that query (access fields, searching type, Boolean logic, etc.) and c) basic technical skills in computing.

At present many users have the essential technical skills to make use, with more or less expertise, of a computer. This number is substantially reduced when it is referred to the conceptual, semantic and syntactic knowledge that is necessary to achieve a moderately satisfactory search. An added difficulty arises in subject searching, as users should concrete their unknown information needs in terms that the information retrieval system can understand. Many researches have focused on unskilled searchers' difficulties to enter an effective query. The mental models influence, users assumption about characteristics, structure, contents and operation of the system they interact with have been analysed (Dillon, 2000; Dimitroff, 2000). Another issue that implies difficulties is vocabulary: how to find the right terms to implement a query and to modify it as the case may be. Terminology and expressions characteristics used in searching (Bates, 1993), the match between user terms and the subject headings from the catalog (Carlyle, 1989; Drabenssttot, 1996; Drabenssttot & Vizine-Goetz, 1994), the incidence of spelling errors (Drabensttot and Weller, 1996; Ferl and Millsap, 1996; Walker and Jones, 1987), users problems
to find alternative terms and to reformulate the query and the movements sequence of the modification (Ferl and Millsap, 1996; Spink and Saracevic, 1997) have been examined.

Results of these researches seem to bring in question the effectiveness of analytic strategies for those end-user queries where the information needs have not been clearly defined. In these cases, improving effectiveness involves creating alternative models in order to reduce the cognitive effort that is required to formulate a search strategy that the information system understands. They represent methods that facilitate some kind of browsing and that imply a dynamic and exploratory searching (Bates, 1989, Marchionini, 1997).

In this context, knowledge structures used traditionally in information systems (classifications, subject headings and thesaurus) have not been considered nowadays as an obstacle, as it has been asserted before by the proponents of automated systems with natural language indexing; instead of this, they have come to be considered as effective tools, combined with advanced techniques of information retrieval (Chan, 2000). Reasons for this new vision are that these tools enhance the conceptual knowledge of users about the information system, its contents and how the information space has been organized, facilitating alternative terms to extend or modify the query (Efthimiadis, 1996). Besides, these established tools are useful for mapping the user's terms to a controlled vocabulary (Bates, 1998, Buckland et al., 1999).

In Spain, empirical studies about end-users subject searching in OPAC's have not been developed. Nevertheless, there are exploratory studies, like the one who analysed the relevance of results from 25 subject searches in CSIC libraries network (González Sereno & Soria González, 1996); another study dealt with a revision of new targets for improving the automated subject access with OPAC's (Seguí i Palou and Vall, 1999). Besides, an implicit problem derives from the non-existence of a Spanish list of subject headings equivalent to Library of Congress Subject Headings or Vedettes-matière of University of Laval. Owing to this circumstance, libraries have fallen back on different methods to fit the subject headings to their own needs, generating both problems of consistency of the representation and the lack of an effective and consistent syndetic structure for providing better subject access in information networks (Gil Leiva, 2001, Jiménez Rodriguez, 1998).

This research work analyses queries in a subject field entered by OPAC end users from a university library, in order to check the match with the terminology used in the indexes of the online catalogs from two libraries which work with different subject headings lists: the Universidad Complutense de Madrid (CISNE) and the Universidad de Sevilla (FAMA). Specifically, this work is aimed at the following objectives:

- to know about the characteristics of searching statements: typology, errors, structure.
- to check the degree of matching between user terminology and the subject headings from both online catalogs.
- to analyse the relationship between characteristics of user queries (MARC fields, number of words, syntactic structure) and the match with both controlled vocabularies.
to examine syndetic structure of the controlled vocabulary, alphabetical subject indexes and hypertextual links of bibliographic records in order to measure their usefulness for retrieving records.

2. Methodology

The subject search statements were extracted using transaction logs dated March to May, 2001 from the Universidad Complutense Library's online catalog (CISNE). The number of queries was 34,908; from this amount were excluded: a) duplicated queries originated at the same day, at the same terminal and with a two minutes interval; b) non legitimate queries - n, r, njijh - ; c) queries in foreign languages - cognitive neuroscience, decision making -. From the remaining number a random sample was collected for a $\alpha = 0.05$ and 95% precision1; overall, the obtained sample consisted of 385 subject search statements, which are the ones analysed in this research work.

In the first place, we examined the characteristics of search statements, such as typology according to MARC fields, number of words per statement, and syntactic structure. For establishing structure categories it all the possible ones were counted, and those categories whose frequency was lower than 1% were included in a unique category called Others.

The next step was to check if there were spelling or typographical errors in the subject queries. Even though it have been established different categories for spelling errors (Drabenstott and Weller, 1996), the types considered in this research work were substitution, transposition, insertion and omission of characters or blanks. These errors were corrected before searching so as to determine what it would be the system answer if there were no errors. The same procedure was applied in the cases of singular/plural variations and correct spelling variations.

At the time of searching, queries were entered in the OPAC Web catalog from the cited institutions, which is managed by INNOPAC Millenium. This system allows to search on rotated subject headings indexes and by keyword matches. In order to search by keyword in subject headings matches was necessary to use REBIUN Web OPAC (Red de Bibliotecas Universitarias), that provides an independent access to each integrated library's collection.

The variables that measure the matching between user expressions and subject headings of the information system are a representation of degrees of similarity, related to terminology and syntax. Although many researchers have undertaken this kind of analysis, the variables that have been used are so heterogeneous that it is complicated to make comparisons fairly. At this point, it was decided to establish categories similar to the ones used by Carlyle (1989), Drabenstott and Vizine-Goetz (1994), Drabenstott (1996) and Drabenstott and Weller (1996); these variables were the closest to our objectives and the most effectively comparable. Despite this, punctuation nor upper/lower case letter variations were considered in this study.

Each user query was searched to see if it fell into the first match category; only when this matching failed it was checked in the next category. Consequently, it is very important to bear in mind the next ordered sequence of categories used in this paper:
1. Subject heading matches.- First of all, to check how similar user vocabulary and syntax are to both controlled vocabularies, two categories of matching were used:
   - Exact match.- Subject heading matches user expression exactly. In this section there are two variants:
     ✓ Word order variation.- Subject heading contains the same words as user expression, but not in the same order.
     ✓ Exact substantive word match.- Subject heading contains the same substantive words which are in the user expression, although word order may differ.
   - Partial match.- Subject headings consists of the same substantive words which are in the user expression, although it includes one or more additional terms.

   We corroborated the relationship between variables such as type of heading, structure and number of words with the variable exact matching. To verify this relationship we used the chi-squared test, significant for p<.005.

2. Alphabetical approach.- In case queries contained words not present in subject headings, the alphabetical index was examined. INNOPAC Millenium system sets user query at the corresponding place in the alphabetical subject index. This option may be useful for user to find an alternative subject heading to continue the search. Based on the first alphabetical index display offered by this system - 12 headings -, we distinguished three possibilities: subject heading may be equivalent, more specific or more generic than user expression.

3. Keyword matches.- At last, those user expressions which did not fall into the preceding categories were searched as keyword in record. We checked the relevance of the first ten retrieved records and verified whether they contained alternative subject headings to lead the user to find relevant records using hypertextual links.

3. Results

The percentage of topical subject searches (MARC tag 650) was higher (86%) than the percentage of the same category obtained by Drabensttot and Vizine-Goetz, 1996 (69.5%); therefore, the rest of fields - 600 (5.7%), 651 (3.9%), 610 (3.12%), 630 (1.3%) - happened in lower proportion than in the cited study. Regarding the number of words by statement, the mean was 2.24 words (ranked from 1 to 8); this figure is slightly higher than the one presented by Drabensttot and Vizine-Goetz (1996) -1.8 words- and by Drabensttot (1996) - 1.6 words-. In relation to number of words, 37.4% of the queries just consisted of a single word, 30.65% consisted of two, 15.06% of three and 17% of more than three. That distribution is directly related to the syntactic structure, as the most common structures are one term (common noun) - 30.13% -, noun and adjective - 19.74% -, and proper noun - 11.43% -, which is usually a surname, or a surname and first name.
<table>
<thead>
<tr>
<th>Structure</th>
<th>Frequency</th>
<th>%</th>
<th>Structure</th>
<th>Frequency</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN</td>
<td>116</td>
<td>30,13</td>
<td>PN+CN</td>
<td>9</td>
<td>2,34</td>
</tr>
<tr>
<td>CN+Adj</td>
<td>76</td>
<td>19,74</td>
<td>CN+PN</td>
<td>7</td>
<td>1,82</td>
</tr>
<tr>
<td>PN</td>
<td>44</td>
<td>11,43</td>
<td>CN+Adj+Adj</td>
<td>5</td>
<td>1,30</td>
</tr>
<tr>
<td>CN+Prep+CN</td>
<td>19</td>
<td>4,94</td>
<td>CN+Conj+CN</td>
<td>5</td>
<td>1,30</td>
</tr>
<tr>
<td>CN+CN</td>
<td>15</td>
<td>3,90</td>
<td>Adjective</td>
<td>5</td>
<td>1,30</td>
</tr>
<tr>
<td>CN+Prep+NC+Adj</td>
<td>11</td>
<td>2,86</td>
<td>CN+Prep+PN</td>
<td>4</td>
<td>1,04</td>
</tr>
<tr>
<td>CN+CN+Adj</td>
<td>10</td>
<td>2,60</td>
<td>Others</td>
<td>50</td>
<td>12,99</td>
</tr>
<tr>
<td>CN+Prep+Art+NC</td>
<td>9</td>
<td>2,34</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Syntactic structure of the queries.

As table 1 shows, the syntactic structure of the statements coincides for the most part with natural language and with subject headings without subdivisions: common noun, common noun and adjective, proper noun, noun and prepositional complement. However, it is observed that some users entered statements not in natural order; this may imply a certain adjustment to subject heading syntax or use of headings proposed by the system - marruecos cultura, juegos educacion fisica, literatura espana -. Due to the differences of categories, comparison with Bates (1989) results is not feasible.

The percentage of searches with spelling or typographical errors (Table 2) was 6.5%; 52% of them were omission errors. This result is higher than the percentage reported in Bleicic et al., 1999 (3.18%), Jones, 1996 (5%) and Drabensttot and Weller, 1996 (5.9%). Singular/plural variations between queries and subject headings represented 3.5% in both universities.

In the analysis of exact matching between user expressions and subject indexes, we found differences for each catalog of the two centres; there are two possible reasons. On one side, users may have adapted their queries depending on the catalog in use; on the other side, user terms may have been derived from used terms in other searches, since transaction logs proceed from Universidad Complutense, where the percentage of exact matching is higher. At Universidad de Sevilla, exact matching – 31.95% - is not very affected by normalization and it is not affected at all by See references. As opposed to this, in Universidad Complutense catalog exact matching increases 3.54% due to normalization and 3.64% due to cross references, obtaining a total percentage of 45.2% for exact matching. In both cases, these results are lower than those reported by Carlyle, 1989 (60%) and Drabensttot and Weller, 1996 (48.5%), but higher than Drabensttot and Vizine-Goetz, 1994 (33%).

Correct spelling variation 2 0,52%
Spelling errors 2 0,52%
Typographical errors
  Substitution 5 1,30%
  Transposition 2 0,52%
  Insertion 5 1,30%
  Omission 13 3,38%

Total of searches with typographical errors 25 6,5%

Table 2. Spelling and typographical errors.

More than 90% of exact matching queries were queries for topical subjects in both catalogs; although this kind of query is the most common, the difference from
the rest of queries was significant \((p<.005)\). However, number of words or syntactic structure are variables with more incidence \((p<.001)\); A fact which speaks for itself is that, in both catalogs, more than 50% of exact matching queries was one word statement, and 30% was two words statements; this relationship was also verified by Markey (1984). In addition, more than 50% of exact matching queries consisted of common noun, and 20% consisted of common noun and adjective.

Table 3 also lists the number and percentage of partial matching: 20.78% in Universidad Complutense and 17.14% in University of Sevilla. This variable comprises those user queries contained in a subject heading that included one or more additional terms (for instance, a geographic subdivision). In these cases, the possibility to retrieve relevant records is reduced; because of it, systems must offer a list of such headings so that it stimulates users to narrow or focus their topics of interest. The obtained figures are higher than Carlyle, 1989 (10%), and also slightly higher than Drabenstott & Vizine-Goetz, 1994 (16.2%) and Drabenstott & Weller, 1996 (14.7%). Exact and partial matches bring users back a list of potentially relevant terms or records, whenever the system allows us to remove non-substantive words and to search by keyword in heading matches or in rotated indexes. But, it is necessary to know what happens to those queries that do not fall into these matching categories; in other words, it is the right time for a subject headings list to prove its effectiveness in searching. And, as table 4 summarizes, the percentage of alternative headings that are provided by alphabetical subject index is higher in both universities than the same percentage provided by normalization. Besides, in Universidad de Sevilla, where exact matching was lower, this figure means a considerable number of queries.

Regarding to searches by keyword matches, 55.11% and 75.58% of these searches provided relevant records in Universidad de Sevilla and Universidad Complutense, respectively. From those retrieved records, over a half in Universidad de Sevilla (56) and about 2 of every 3 in Universidad Complutense (73) contained at least one alternative heading useful to continue the search or to browse the catalog with hypertextual links. This figures represent 15% (Sevilla) and 19% (Complutense) of the total of queries statements. Summing up, if we add these percentages to the results obtained in the preceding category (alphabetical approach), we find that possibilities to find relevant records with the help of browsing through system tools (links between records indexed with the same headings) and alphabetical subject heading index account for more than 25% in both university online catalogs.

<table>
<thead>
<tr>
<th>Univ. of Sevilla</th>
<th>Univ. Complutense</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exact matching</strong></td>
<td><strong>Frequency</strong></td>
</tr>
<tr>
<td>Direct</td>
<td>115</td>
</tr>
<tr>
<td>See reference</td>
<td></td>
</tr>
<tr>
<td>Word order variation</td>
<td>2</td>
</tr>
<tr>
<td>Exact substantive word match</td>
<td>6</td>
</tr>
<tr>
<td><strong>Total of exact matching</strong></td>
<td>123</td>
</tr>
<tr>
<td><strong>Partial matching</strong></td>
<td>80</td>
</tr>
</tbody>
</table>

Table 3. Exact and partial matching.
4. Conclusion

In the matter of characteristics of queries, we found an extensive range of structures, but the most common was topical subject and with only a few words. Results analysed above show that queries with more probability to be exact matches are those with simple structure and few words. The percentage of queries with typographical or spelling errors or singular/plural variations accounted for 11%; to solve this problem, a system of detection of misspelled words (Drabensttott and Weller, 1996; Walker and Jones, 1987) could be implemented.

To improve the match between user statements and controlled vocabularies it would be expedient to incorporate into online catalogs new techniques of information retrieval; these should be combined and promoted in a way like the search trees proposed by Drabensttott (1996). It has been observed above that subject heading lists are a very valuable tool for enhancing access to information: an improved syndetic structure and crossed references provides alternative terms and increases the proportion of exact matches. Besides, linking records which are indexed with the same subject headings brings users the possibility of effective browsing.

Alphabetic subject heading indexes may become a important instrument for guiding user through the library collection, focusing or narrowing users queries, as Ariadne’s thread guided Theseus to find the way out of the maze. That is the essential aim of their knowledge structure, which can and should be developed by technology (hypertext had opened news ways for retrieval) and by consolidated indexation instruments (alphabetical indexes and references).

Notes
1.- To calculate the sample size it was used GRAMNO 4.0, and to select queries randomly and the rest of statistical process it was used SPSS 10.0.

References.


Linda Banwell
Information Management Research Institute, School of Information Studies, Northumbria University, Newcastle upon Tyne, UK

Developing an Evaluation Framework for a Supranational Digital Library

Abstract: The paper will explore the issues surrounding the development of an evaluation framework for a supranational digital library system, as seen through the TEL (The European Library) project. It will describe work on the project to date, and seek to establish what are the key drivers, priorities and barriers encountered, in developing such a framework.

TEL is being funded by the EU as an Accompanying Measure in the IST program. Its main focus is on consensus building, and also includes preparatory technical work to develop testbeds, which will gauge to what extent interoperability is achievable.

In order for TEL to take its place as a major Information Society initiative of the EU, it needs to be closely attuned to the needs, expectations and realities of its user communities, which comprise the citizens of the project’s national partners. To this end the evaluation framework described in this paper, is being developed by establishing the users’ viewpoints and priorities in relation to the key project themes. A summary of the issues to be used in the baseline, and to be expanded upon in the paper, follows:

1. Establishing the differing contexts of the national library partners, and the differing national priorities which will impact on TEL
2. Exploring the differing expectations relating to building and using the hybrid library
3. Exploring the differing expectations relating to TEL. TEL needs to add value – what does this mean in each partner state, and for the individuals within them?

1. Introduction to TEL

TEL (The European Library) is a thirty month project, funded by the European Commission as part of its Fifth Framework Programme for research. It aims to set up a co-operative framework for access to the major national, mainly digital, collections in European national libraries. TEL is funded as an Accompanying Measure, designed to support the work of the IST (Information Society Technologies) Programme on the development of access to cultural and scientific knowledge. TEL will stop short of becoming a live service during the lifetime of the project, and is focused on ensuring co-operative and concerted approaches to technical and business issues associated with large-scale content development. It will lay the policy and technical groundwork towards a pan European digital library based on distributed digital collections, and providing seamless access to the digital resources of major European national libraries. It began in February, 2001, and has eight national library partners: Finland, Germany, Italy, the Netherlands, Portugal, Slovenia, Switzerland and the United Kingdom. It is also seeking to encourage the participation of all European national libraries in due course.
Included in the project plan is the objective of developing an evaluation framework for TEL, as a benchmarking baseline to be implemented by the partners once TEL becomes a live system.

The TEL project is being evaluated at 3 levels:

- First level: completion of Workpackage deliverables, to be monitored by the TEL Project Manager
- Second level: overall project performance in relation to the IST programme priorities
- Third level: establishing user viewpoints as the basis of an evaluation framework to be used as a benchmarking baseline for the eventual live TEL service.

The second and third levels of evaluation are being undertaken by an independent evaluator, and it is with these that this paper is concerned.

The project activity is being pursued through six Workpackages, focusing on:

WP1 – Publisher relations. The consortium is working with publishers of electronic materials and publisher organisations to establish co-operative approaches to business, licensing and copyright matters.

WP2 – Business plans and models, designed to maximise the benefits of cooperation through the development of joint or individual business plans and models ready for implementation in the operational phase of TEL.

WP3 – Metadata development, aiming to develop a concerted best practice approach to metadata standards and schemas to support access to digital material. The agreed approaches will be tested in WP4.

WP4 – Interoperability testbeds. This WP will carry out preparatory work prior to the development of the operational service. It will develop a functional specification, and benchmarks will be defined from the outputs of WP1-3. There will be 2 testbeds, focused on Z39.50 and XML. Testing will include scalability of access and multi-lingual capability.

WP5 – Dissemination

WP6 – Management aspects, which includes the evaluation activity.

2. Developing the evaluation of TEL

Evaluation activity is concerned with arrangements which capture, systematise, disseminate and act upon project experiences (Tavistock Institute, 1996). A review of existing literature on evaluation and examples of its use in other digital library projects, was conducted at the start of TEL. The Tavistock Institute (1996) has produced guidelines for the evaluation of digital library projects, and both Saracevic (1999) and Marchionini (1999) write about specific project evaluations. Work on the return-on-investment model (Collis and Moonen, 2001) provides the broad framework by which partners could take forward the evaluation framework for use in a live TEL service beyond the scope of the current project. Related work by the author on other projects (Banwell et al, 1999, 2001) has also aided the development of the TEL evaluation. The European Commission’s guidelines (2000) on its evaluation of project proposals, was useful in indicating themes of current EU concern.
The broad context of TEL as an EC funded project provides the basis for the evaluation framework described in this paper. There are several aspects, which provide a starting point.

- the **Information Society Programme themes**, for example, the "informed citizen" is a prominent theme – do TEL partners have a profile in this context, and what is their approach to serving the general public? Has community added value and contributions to community social objectives been achieved? Have community economic development and dissemination criteria been met?

- the requirements associated with being an **Accompanying Measure**: the project evaluation, which would be suited to a full RTD project, is not appropriate for TEL, which will ultimately need to incorporate the evaluation of an on-going service. The evaluation of the TEL project phase therefore needs to lay the foundation for an eventual service evaluation.

- at **partner level**, to discuss the added value to be had from participating in TEL for the library and its users, for example, reviewing the extent to which TEL is encouraging partners to develop resources that, in co-operation, are greater than any would have on its own

- **common questions** to be asked when evaluating a europroject are about the short to medium term outputs of the project, the learning gained and the extent of its diffusion leading to the establishment of new practices which might act the catalyst for further change.

A variety of activities will build up different layers of information in support of the evaluation. In the first year of the project, meetings were held with in the project team in order to develop the framework for the evaluation. These were followed by broader based meetings between partners in the project, and meetings with the wider stakeholder group, at the TEL Milestone conference held in Frankfurt in April, 2002.

There will be two outputs from the evaluation and monitoring activity:

- The overall, **summative project performance review**. This will be end-of-project activity, which will include synthesis of Workpackage level evaluations and other documentation generated in the project, and of data collected direct from partners and partner communities.

- **Formative, monitoring and evaluation activity** designed to establish and monitor TEL user viewpoints, link back in to the evaluation of specific Workpackages in an on-going way, and inform the development of a benchmarking baseline for TEL.

The TEL evaluation is essentially user centred, where qualitative data are as important as quantitative data. It is multifaceted and multilayered, as described above, with data triangulated to enhance its trustworthiness and objectivity. Formative evaluation gives a longitudinal character to the evaluation, whereby it becomes problem solving, feeding back into the project. The approach aims to be holistic, naturalistic. It is taking place in an uncertain environment: the future shape of digital libraries is unknown, meaning that the evaluation must provide the partners with a flexible and adaptable tool for their future use. It is evaluation with a research focus - it seeks both to explain the specific digital library, and relates it to wider issues (Marchionini,1999).

The evaluation process is progressing through the following stages:
Establish a written evaluation plan, taking into account EU project evaluation requirements, and building on best practice in the field. The plan was delivered in September, 2001, and included in documentation for the first six monthly review by the EU. The feedback received stressed the importance to TEL of finding out in detail about its potential users, thereby reinforcing the approach being taken in the TEL evaluation.

Documentary analysis of the detail in the scoping exercises carried out by all Workpackages, provided a detailed starting point for the evaluation. The market research activity in WP2 has been especially useful in providing information on user contexts, views and service evaluation criteria.

User panels were planned in all partner libraries to represent as many views as possible from within the national user communities, and as an important mechanism for on-going contact with users and potential users of TEL.

Formative evaluation was planned through structured six monthly contact with users and potential users, either electronically, or face-to-face at Workpackage meetings, and at conferences (such as the TEL Milestone conference in April, 2002) and workshops. Questions seek to identify problems and solutions in the development of TEL in the broad areas of content (issues here will be accessibility, accuracy, richness), technical infrastructure (e.g. speed, ease of use), and user perceptions of the role, importance and potential impact of TEL to users as individuals, and nationally. Shared understandings are being sought through investigation of variations at national and individual levels.

Formative evaluation activity is reported through the preparation of six monthly reports for the TEL project manager. It will also feed into the summative evaluation report at the end of the project, with appropriate dissemination.

The TEL scoping studies carried out to date show that the current situation with regard to users and use of digital resources in TEL partner libraries is largely a continuation of the pre-digital situation. The current emphasis of collections and services is on national language, literature and culture, used largely by higher level students, academics and some personal researchers/authors. Most partners have few or no statistics about their current users and usage. Where data are available they are for traditional reading room registrations. No partners have data on their internet users. There is a problem with the comparability of even what data are available – definitions and the recording of statistics vary between partners, and the tables below can be viewed as indicative only of current situations. Nevertheless, variations in practice between partners are suggested and illustrated in the following figures, which provide an illuminating baseline for the project.

In addition, three partner libraries provided figures which show that around 80% of the reading room/registration statistics were by students and academics, and most of the rest were by 'other professions'.

Through the user panels, potential users of TEL are being asked questions to establish their current information behaviour, especially in relation to digital resources, and for their views on the added value that TEL might bring them.
<table>
<thead>
<tr>
<th>Type of digital resource</th>
<th>Lib 1</th>
<th>Lib 2</th>
<th>Lib 3</th>
<th>Lib 4</th>
<th>Lib 5</th>
<th>Lib 6</th>
<th>Lib 7</th>
<th>Lib 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>scientific jnl articles</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>X</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>newspapers</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>manuscripts</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>digitised books</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOC's books</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>incunabula</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>early printed books</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dissertations</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>maps</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>music scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grey literature</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>official publications</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ephemera</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sound-wildlife</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sound-dialects</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sound-ethnography</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>illustrations</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>portraits</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>posters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>photographs</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>images</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>unspecified</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1 TEL partner digital resources (current and planned)

Partner libraries are being asked questions to elicit responses which map on to EU priority concerns and which, by the end of the project, will be able to answer questions to inform the overall project review, such as:

➢ To what extent has a centralised management process been created for the integrated process of search, locate, order, receive and pay?
➢ To what extent have centrally negotiated licences for the delivery of material been achieved?
➢ How much research has happened on non-native language needs?
➢ Does the TEL business plan make recommendations which would add value through TEL? Does it indicate the issues of transition from project to operational service? Does it have an outline marketing plan?
<table>
<thead>
<tr>
<th>User category</th>
<th>Lib 1</th>
<th>Lib 2</th>
<th>Lib 3</th>
<th>Lib 4</th>
<th>Lib 5</th>
<th>Lib 6</th>
<th>Lib 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Higher education students</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Academics/researchers</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Other professions</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Personal research/authors</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Senior citizens</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Primary pupils</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Secondary pupils</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lifelong learners</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Booktrade</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other firms</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Libraries, museums</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Local</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>By discipline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Humanities</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Law</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Economics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Business</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Social sciences</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medicine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

Fig. 2. TEL partner user categories

➢ To what extent have common prices and performance measures been agreed between the partners?
➢ Has a strategy to address non-use been developed in each partner library?

3. Conclusions

A number of issues for resolution have already been raised in relation to the TEL evaluation:

➢ About the users - who and what are they? How to access them in a comprehensive and representative way? Responses from partner countries have varied widely.

➢ About culture, language and meaning - making allowance in an evaluation for such wide variations between partner customs and practices, can make comparisons difficult.

➢ About user panels - setting up and establishing communication with user panels has proved impossible in some partner libraries, making it impossible for the
voice of at least part of the TEL user community to be heard. Identifying the wider user constituency is itself a difficulty.

- **About formative evaluation** - it occurs during the life of a project and needs to be fed in, in a meaningful way, which is complex.
- **About evaluating digital libraries** - they are a moving endpoint, making the development of benchmarks inappropriate at the present time.
- **About moving forward on a very short timescale** - the project lasts 30 months only, and is large and complex.
- **About TEL evaluation being very small scale** - it represents 8 weeks effort over the whole of the project, making it a problem to collect enough and the right data in a short time.
- **About evaluating an Accompanying Measure** rather than an RTD project. The EU has stressed the importance of the involvement of users. But it is difficult to engage potential users of a service, which will not be operational during the lifetime of the project.
- **About the evaluation type** - is it service or project evaluation? The aim and nature of the evaluation differs between the two types. For TEL, the evaluation is of both types: summative for the project and formative for the proposed service, making it a complex activity within its limited budget.
- **Genuine user-based evaluation is a low priority in EU projects** - evaluation is generally seen as being a rather mechanistic, tick box, checklist type of activity. A more broadly user-based evaluation is unexpected and therefore can be viewed with some scepticism.

Despite the difficulties raised in the points listed above it is hoped that by doing and disseminating the user-based evaluation activity in TEL, it will thereby be possible to generally raise the profile of user evaluation in europrojects. But, as in all supranational projects, it is very difficult, and it should be larger scale and of higher priority for the EU, if their community-wide objectives are to be met. It is hard, immersed at the present time in doing the evaluation, not to see the activity merely in terms of barriers, but it is hoped that it will ultimately succeed in laying a foundation for the future service of TEL.

References


Banwell, Linda, Day, Joan and Ray, Kathryn (1999) **Managing Organisational Change in the Hybrid Library. report to UKOLN.** Department of Information and Library Management, University of Northumbria

Collis, Betty and Moonen, Jof (2001) **Flexible learning in a digital world. Experiences and expectations.** Kogan Page


Antonio García Gutiérrez  
Department of Journalism, University of Sevilla, Sevilla, Spain

Knowledge Organization from a “Culture of the Border”: Towards a Transcultural Ethics of Mediation

Abstract: The social construction of the digital memory, let us call it “exomemory”, has traditionally been a task related to aseptic procedures and tools but, in fact, it is an activity crossed by complexity and mediation. The positivist model claims for objectivity as the frame and goal in and for which electronic and external memory workers and thinkers have to fight and strive. The theoretical concept of multiculturalism is a dangerous slogan and not sufficiently critical as to tackle the rights of diversity and singularity even within a given (but not real) “monocultural society”. Exomemory mediators as librarians, archivists, documentalists or virtual curators are not capable of addressing their tasks from a holistic approach compatible with every culture without determining their products and services of symbolic value from an hegemonic position, should it be at local, national or global level. So, these professionals and scholars have to practice reflexivity and include other metatheoretical concepts in their ordinary actions so that users may know who is behind the analysis, “whose are the tracks”. To achieve this aim, the field of research called “Knowledge organization” must be opened to a new paradigm in which Critical Theory and Hermeneutics go together. Several theoretical and metaphorical terms commonly used are reviewed and forced to their paradoxical limits. The essay stands for a “culture of the border” as the best imaginary place to depict and accept those contradictions denied by dogmatic and hermetic intelligence.

1. Introduction

Let us think, as a starting point in the mood of this theoretical essay, of what I understand as a “culture of the border”: Let us take the ambiguous prospect, privileged by this metaphor, that invites us to observe and describe the whirlwind of a world in permanent mutation — too quick or too slow to be completely perceived by our rude senses, as it was stated by Nietzsche—, either visible or imagined, internal or public, objects or phenomena, from hybrid and forever-changing positions. The border, as a complex place —in terms of the epistemology of Edgar Morin (1996) and Boaventura Santos (1989)- favours the adoption of new and coherent logics about what a limit means. A boundary is, in the end, both a noological (non-physical) body and a “chrono-topos” (time and space entity). In that territory, we have to take into consideration, evenly and in its full integrity, the connecting points as well as the “rupture lines” (Deleuze and Guattari, 1994). The limit means here a brand new meeting between the same old enemies, established as opponents by the positivist paradigm: closure and opening, rigidity and flexibility towards an external and alien environment to which we are also external and foreigners, in the middle of a never-ending identity that only a theory admitting the existence of closed systems is capable to deny.
The closed system conception is protected by a demarcationist logic constitutive of the modern sciences' paradigm, whose effects may be verified in our elemental and ordinary way of approaching, working, possessing or thinking of an individually and socially constructed world that, even as a mere construction, is not mutation-free. The world, that we know and we are, is just the perception of a devastating inter-action (or else, a trans-action) that desintegrates everything so that everything may fortuitously re-integrates itself into new entities. We have to bear in mind too, as a part of the concept of demarcationism, an old practice of power and a fight already present in the evolutive spring of all species. This unstoppable force of nature has also reached human reason diverting it towards rationalization, as a unique/exclusive mode of perception, already in the path of a mental illness (Morin, 1996).

The history of the organization of knowledge has been the history of a process of rationalization of knowledge that has forced other real and modal ways of cognition to the exile, ruling out, for instance, the actuality of an irrational mind which cannot be simply excluded either from participation in our representation of nature or from decision-making. Obviously, all this affects the construction and organization of knowledge in the electronic environment, what I would like to call now “digital exomemory” (García Gutiérrez, 2002) a space which has not yet been fully demarcated but, through a step-by-step policy of occupation, it is being reduced by the same old forces and aims of domination and colonialism which have re-adapted to these new digital territories.

Demarcation, as a human practice in the field of knowledge, is a useless utopian approach to reality when it systematically intends ontological cleansing. Nonetheless, if “demarcationism”, as a doctrine (teleologism) or just as a simple goal, is based on the simplification of the difference (and not in search of singularities), “to demarcate” is not itself an intrinsically negative practice: what we have to take into account is the code and intentionality of the demarcation forces and the symbolic imagery that could be displaced, regenerated or wiped out.

Our field of knowledge, by the way, “knows” a lot about lineal classification and conceptual cleansing. However the wake of reality and the cognition world “organize” and shape themselves in a rhizomatic and paradoxical “order”. For this reason, I am committed to the new emergent paradigm based on a doubly critical-hermeneutical approach, in taking the theory and practice of classification and knowledge organization, biased by positivist model, to their limits, so that to provoke the crisis of our intelligibility matrix right in the absurd of the border. So, after the deconstruction of a text, as solid as three centuries of positivist sediments and important success (also accompanied by certain mental and sociocultural dystrophies), I take up the opposite route towards the reconstruction of “integral sense”, that comes, paradoxically again, from de-classification and disorder. This implies seeking other possible organizational orders of the world as well as admitting the disorder itself as a possible form of order from a caologist version of the concept (Prigogine, 2000). Some of the traits of the new thinking, depicted here, are constituent features of what I would like to introduce as “Interactive Epistemography” (García Gutiérrez, 1998, 2002), a transdisciplinary configuration that sets up, among other principles, the search of opened-horizontal semantic networks counteracting (against) human and political imposition of verticalism and techno-cultural hierarchies. The tree, as a botanic phoroi since Empedocles,
“naïvely” conveys structures, moulded on scientific and political power, to the organization of the digital exomemory.

The demarcationist rule operates in the knowledge organizer’s subconsciousness as a guideline notion of the “otletian” positivist universe. Its well-known maxim, “a single place for each thing and a single thing in each place”, illustrates that reduction of diversity. Time has come to balance our, so far unavoidable, segregationist and decontextualist task restoring, in the core of it, a consciousness of the re-interpretation role that we all play, as both mediators and end-users, in order to get an epistemological framework whose ethos should be the bluring of borders, fading away those imaginary walls raised on behalf of discrimination and purification. In their place, linking-concepts, based on a crossbred status, instead of fencing-concepts, should be located.

Let us have a look at some preliminary considerations about the formal collision of knowledge organization with other alter-ego constellations (that is, following hegelian synthesis, a sort of external entity to our construction, dwelling, at the same time, in the heart of it): discourse, memory and culture:

1. As to the operation, knowledge organization not only deals with organizing, as it is well-known, but also addresses selection schemes, analysis, transcode procedures, representation and access supplies. This last expression should be replaced by the concept of participation in order to remember democracy and the public property of knowledge on a global scale rather than private property and market distribution. These operations, far from being neutral, banal or aseptical, are, all of them, deeply mediated.

2. As to the application field: we are not just coping with knowledge itself (this is a metaphor, too) as an equivalent to recorded wisdom or exomemory but also with any raw inscription or trace over which a discourse has already been elaborated comprising a diversity of predetermined structures and conditions existing at any other discourse production level: human beings not only file what they know but what they feel, imagine, remember or dream of, too. That is the reason why we have to overcome the reduction of the concept of knowledge, bringing to our practices an open debate on the co-existence de facto of epistemical, doxological and emotional elements in a threefold level: text producers, users (participants) and mediators (epistemographers).

This new rhetorical conception of scientific knowledge includes incursion-studies on common sense, psychagogics (ancient aristotelian concept related to non-rational argumentation directed to persuasion), emotions and, obviously, on rational arguments themselves when pointed at persuasion which is the goal of every text and statement, obviously including those of exomemory mediators (vid “sense communities” in Berger and Luckmann, 1998). Thus, information is not free from emotion and rhetoric. Emotions and passions are transferred ergo they are also informations.

To reflect on possible organizations of our external memory, we have to combine several angles of observation relating to a general theory of mediation, both in terms of Michel Foucault’s suspicions hermeneutics and from the ideological analysis drawn from the perspective of the School of Frankfurt, namely Jürgen Habermas (1999) and Otto Apel (1996) (and their call for the need of a dialogical ethics seeking consensus). Nevertheless, other fields addressing meaning and sense from individual, social and cultural points of view must imperatively be
run: symbolic interactionism, etnomethodology, knowledge sociology or social psychology among others.

So, our theoretical task is committed to a social perspective of the procedures and tools but at the same time any proposal has to be workable and feasible as an industrial activity which is not incompatible with the care of sense, diversity, plurality and social transformation. This means to introduce esthesia (Sodré, 1998)—rather than anaesthesia—in the industrial process of knowledge itself.

2. From multiculturalism to transculture

Let us analyse, now, the concept of multiculturalism under the suspicion of a double-edge. If a culture is an open system, how can anybody demarcate it? The expression “multiculturalism” uses a prefix, “multi-” which is acceptable when it means variety although it may prevail a mere sense of quantity over a complex process, the culture, that impedes any enumeration and even hinders proper naming. Do we know the external and internal limits of the Andalusian, Spanish or European culture? Can we inventory and purify cultural taxonomies sharply underlying their components and borders? If the answer is no, then we are in the presence of an abstract notion—multiculturalism—that does not deserve a deeper ethical discussion.

To understand the concept of multiculturalism we should previously try to explain what a culture is, which is the main drawback. Culture rejects single and simple theoretical approaches without determining the concepts taking part in the definition (Jameson and Zizek, 1998; Olivé, 1999). Every epistemic subject is beforehand a cultural actor. This obstacle hampers a single conception of culture and even more a single classification of cultural typologies. The theory of culture shares this “setback” together with the theory of discourse: Both of them only admit empirical, biased, intentional and simplified typologies, therefore, hardly sustainable.

The knowledge organizator is also trapped by these inconveniences. That is why, rather than simplify cultural realities, scholars and practitioners have to work together to build a transcultural ethics which gives legitimate support to the social construction of the exomemory, using categories transculturally acceptable, and a body of rules helping to detect racial, ethnic, gender or any other prejudices of the like. This deontological code should be linked to an ad hoc propedeutics of mediation that do not need to deny mediation, as a real presence, but must foresee a way of making it explicit and prepare wide-awake participants.

Another big problem for epistemography is the propensity to convert cultural stereotypes in formal categories of organization, for the seizure of a set of cultural traits, from a given group or society "culturally pure", is distorted by the redundant use of topics and stereotypes in a sort of categorial rank defined in an imaginary "cultural catechism" brought up as a single photography of a cultural reality complex enough and fast to be frozen by cameras, intentions or theories. So, any reliable approach to that changing world has to be as nimble as reality is.

If, in the end, we manage to simplify a culture, or a well-bordered set of cultures, that is multiculturalism, if we tackle all that stuff as steady entities permitting it fragmentation into independent and parallel subsets (even under the suspicious goal of preservation) we will get well-limited and isolated cultural
frames that have nothing to do with the real world. And we may also get an even worse result: the isolation of a culture, based on the emphasis of different features, could lead some political projects to "mummify" their cultural realities in order to both export them as simple commercial stereotypes of the country or prevent them from foreign pollution, in the line of current ultra-nationalistic doctrines. Under the pretext of preserving a local culture from alien invasions, or from a single global culture, there could be the risk of turning that frozen culture into a single compulsory local scheme for all the citizens living in a given territory.

That is the reason why I only accept those multiculturalist theories that consider a culture as a dialogic, interactive and wide-open system fed and grown from actual and current interaction and not only by means of traditions and nostalgias, most of them constructed or reconstructed a short time ago or to the convenience or extracultural interests. A democratic and open cultural system must strengthen its actors with a high level of education on democratic, human rights, and local values. But these measures must respect and also protect the right to choose any other cultural preferences either radically different, hypocultural or even an anomic behaviour. The only acceptable cultural policy is that favouring heteroculturalism, i.e. that policy which lives up to the human right to singularity. That principle must also be conveyed to any categorial scheme of knowledge organization. This general position taken so far is what can be understood as transculturalism, a transversal look that has to be placed in the contradictory and complex heart of metacognitive demarcation, that is, in classification of human knowledge and memory.

3. Prejudices under indictment: Towards a transcultural ethics of knowledge mediation

Finally, I have to bring up some ballast and ingrained bad habits, so far inseparable from our research and practices. Knowledge organization researchers and practitioners should be aware of them in order to detach them from exomemory construction:

1. The principle of hierarchy based on the porfirian tree, as a dualist scheme and plan of reality, has to be replaced by a rhizomatic conception of the world. This conception not only sees every single element, as a product of previous interaction process, but the interaction itself.

2. Demarcationist logics, grounded on the principle of exclusion instead of assuming and striving differences as an inclusive force, is active in all levels and status of operation and theory. Disjunction has to be overcome by consensus measures leading to the rapprochement of far and opposite ends. These radical (from the common roots) approaches must protect diversity and singularity as individual instances resuming and harmonizing concrete and holistic perceptions. Positivist demarcation programme generates fragmentation, decontextualization and hyper-specialization, that is, drastic interdisciplinary disconnection. Fragmentation opens up the door to free manipulation by extracientific entities and interests.

3. Positivist demarcation is also responsible for the exclusion, from an authorized single model of cognition, of other ways of perception and construction of knowledge and reality: from Philosophy to Metaphysics (detached from
material Physics since Galileo and Newton), from Humanities to common sense. Positivism accused prior integrated ways of cognition of creating polluted knowledge. In order to cleansing knowledge, positivist dogmatism ruled out observers from observation imposing, for this purpose, the directives of natural sciences over social and human sciences. At present, however, even the new Physics itself deny full neutrality, since Heisenberg, Gödel, Bohm or Prigogine.

The positivist doctrine has established a conception of the world based, to my mind, on wrong values as exclusion, competitiveness, exploitation, functionality, success and merchandising. These values have swamped our societies, mentalities and daily practices. Our field of research must get rid of those élite habits, most of them imposed by a global policy of market interests, consumption fostering, and a non-aseptical technical language that stands out for concepts such as Quality, Efficacy, Welfare or Innovation, new totems expanded as indicators of modernity and progress (also of new ethics?). All these concepts should be reformulated mainly from social and critical perspective.

We have to turn towards a new social framework of epistemological, ethical and technical principles in knowledge organization, summed up now as deontic statements:

- No discourse will prevail over the others (natural sciences over social sciences, social sciences over humanities and culture, élite culture over popular culture, popular culture over mass culture, etc.)
- No cognition process will prevail over the others (including emotional and persuasive discourses) unless the organization of a specific application field needs it (scientific discourse will need an overall logical and categorial rationality, a given culture will privilege, on the contrary, emotion, tradition and affection).
- No culture will prevail over the others. No cultural category or feature is superior to its neighbour’s. There is neither a Great Culture and auxiliary ones, nor the big science and the minor sciences, nor the perfect discourse and the secondary discourses that have to imitate the model.
- No users will be privileged over the others (scientists, politicians, journalists, lawyers) unless required again by a specific application. There is no great modelic user and the grey audience. Whatever the application, no knowledge system should ignore the existence of a general and transversal user.
- No system has to ignore its antigravitational and leaking forces for these centrifugal lines set up points of prolific intersystemic linkage. Every entity in a system has also its centrifugal forces providing more intrasystemic coherence.

References
Abril, Gonzalo (1997): Teoría general de la información. Datos, relatos y ritos. – Madrid: Catedra
García Gutiérrez, Antonio (1998): Principios de lenguaje epistemográfico. — Sevilla; Granada: Instituto andaluz del Patrimonio histórico; Comares
Giddens, Anthony; Habermas, Jürgen; Jay, Martin; McCarthy, Thomas; Rorty, Richard; Wellmer, Albrecht y Whitebook, Joel (1994): Habermas y la modernidad. — 3ª ed. — Madrid: Cátedra
Morin, Edgar (1996): Introducción al pensamiento complejo. — Barcelona: Gedisa
Sodré, Muniz (1998): Reinventando la cultura. La comunicación y sus productos. — Barcelona: Gedisa
The Importance of Context in Resolving of Conflicts when Sharing User Profiles

Abstract: The Web has transformed our way of viewing information retrieval. Much research work is focussed on intelligent information retrieval and personalised presentation of information based on user's preferences, interests, etc. However, one aspect which has been neglected thus far is the treatment of user profiles. Currently user profiles are assembled by information providers with no regard to future sharing. However, in the future there is likely to be growing need to share user profile information. This paper discusses some of the problems involved. A prototype has been built to explore these ideas.

1. Introduction

The exponential growth in the amount of electronic information available to people, especially through the Web, is leading to increasing problems for search processes due to the sheer volume of information available. This has led to interest in more intelligent support for the user, including:

- Intelligent filtering, where the challenge lies in developing an information retrieval system that uses data about the user to filter the information retrieved and return only relevant solutions.
- Customisation, where the challenge lies in ordering and other presentation aspects to maximise the effectiveness to the user of the filtered information.

These two facets, filtering and customisation, are generally referred to as personalisation, which is still in its infancy (Oard, 1997). It is assumed that the basis of personalisation is the user profile that holds knowledge about the user, such as the user’s preferences, interests, etc. The current state of user profile development is still in a very early stage, and there is little agreement about how these should be constructed. Profiles tend to be constructed for and used within a narrow range of services, for example, favourite categories of movies. To increase the level of personalisation requires either user profiles to be constructed for a broader range of services or the sharing of a number of separate user profiles to try to gain the necessary scope of knowledge about the user. This paper is concerned with the second solution. Current approaches used are ad hoc, with each information provider building user profiles without any regard to potential sharing of information in the future (Fink and Kobsa, 2000).

This has certain parallels with the field of databases. Initially, databases were developed to solve specific information storage problems, with no thought given to the problem of how to integrate them. With time it has become apparent that for many applications users need to access information from more than one database, and thus it would be helpful to link sources together to provide integrated access to them as a single source. Solutions have focussed on integrating heterogeneous information from distributed sources (Austin, et al 1994).
The sharing of user profile information is likely to follow a similar path, with initial isolated ad hoc systems, moving to the sharing of profiles in the future. This paper explores some of the problems. The next section discusses issues involved while section 3 expands on the specific issue of user context. Section 4 discusses some of the problems with using context, which are currently being studied by a prototype.

1. The Issues

As noted in (Motomura, et al 2000) "Unfortunately, most user models are constructed in a specialized form that is not applied to other systems or domains. This specialization makes difficulty in sharing user models as common resources for developing information retrieval systems and for researching cognitive characteristics in various users." In considering the differences between information stored in different user profiles, there are three important aspects to take into account:

- How the information is represented in the profile
- What the information means
- The context of the information

1.1. Representation

Each user profile is constructed to work with a specific application and as a result the representation chosen will be determined by the application. The choice of attributes, data types, data structures, will vary from one application to another. As a result two user profiles may contain identical information and yet appear to be very different. This is the same as the problem of integrating heterogeneous database systems (El-Khatib, et al 2000), and solutions to the latter problem may well be applied to the user profile problem. For example, in the case of databases, metadata may be used to describe the format of data in individual data sources. This is then used to resolve the heterogeneities in data from different sources. The problem of representation will not be dealt with in this paper.

1.2. Meaning

The meaning associated with any item of information is dependent on the domain from which it is drawn. Where domains are totally different, sharing user profiles is unlikely to be useful, e.g. in the movies and travel domains. However, across domains we can find similar attributes appearing, for example wheelchair access. This attribute is relevant to a large number of different domains, e.g. restaurants, travel, hotels, etc. This is also an example of an absolute attribute i.e. a person who is confined to a wheelchair is likely to be confined to a wheelchair regardless of the domain. In most cases, attributes are not absolute across domains the value being dependent on the domain.

1.3. Context

There has been significant interest in different aspects of context and various definitions have been put forward (Brezilion, 1999, Turner, et al 2001). However, there is no generally accepted definition of the term and to some extent it depends on the application. In this paper, the term will be taken to refer to additional information associated with an event or object, which forms the setting of the event or object and helps to determine its true meaning. For example, each time a user visits a restaurant, she is not going to have the same preferences. Parameters such
as, who she is eating out with, where she is, when she wants to eat, or why she wants a restaurant in the first place, will all influence the set of user profile attributes. It is to these parameters that we apply the term *context*. It is the context that applied when the user profile attributes were captured that influences how the profile may be used when answering a query. Depending on the context, different answers may be returned.

2. Context

The definition of context extends to cover many parts of the information retrieval process. Each user profile attribute has a context that reflects the circumstances under which the information was captured. This includes both the user's circumstances and the information provider's approach in gathering it. The result that is retrieved in answer to a user query has a context. Each user query has a context, which may not be explicitly stated. There is also the context of the user’s device that describes things that are independent of the user profile.

2.1. User Profile Context vs. Query Context

The *context of the user profile* may be inferred from the user's previous queries. When the profile is generated from analysing past user behaviour this relationship is direct. As the profile changes in response to queries, these changes have the same context as the query that caused the change. As a result, it is possible that different attributes in the same profile may have different contexts. When the profile is generated from the user entering a set of preferences, this too has context but this time the relationship with the query is reversed. Here the user is supplying information for a particular context of query. In this case, it is possible that the user may believe that the information has one context while the collectors of the information assume another. If the context can be deduced reliably then this will improve the use of user profile attributes for personalisation. One solution would be to include the context information as part of the user profile.

It is the *context of the current user query* which will determine the user profile attributes that can be used to answer the query. Hence, the context of the user query must be known or deduced in order to select profile attributes. This deduction may not always be possible but with enough information certain contexts can be eliminated. If the user wants to book a table for one, they are clearly on their own (see 3.2.2). Likewise, it may be possible to deduce whether they are either on holiday or on business. If the query has enough information then, using simple deductions like this, it will be possible to reduce the number of contexts. If necessary, further clarification of the context can be sought from the user directly. In some situations, for example where a history of previous user queries is kept, it may be possible to deduce the context of the user query from their previous queries (Mathews, et al 2000).

2.2. Classification of context

The user query is the root source of user context. Thus, the information implied or explicitly stated in the query suggests different types of context. What is needed is to look for triggers in the query that indicate a context. Of course, there may be a context conflict where the user may believe that the query has one context while the system implies another. What follows is a first attempt at classifying context.
2.2.1. The Context of Ownership

A user may enter a query on behalf of someone else. For example, if a user buys a gift, and this information ends up being captured in the profile, profile attributes resulting from this action are not necessarily the user's own preferences. Thus, although the profile may be owned by an individual, not every attribute of this profile may be applicable to the owner. What has been captured is not the user's preferences but what that user believes about another. It is clearly information that should not be used to personalise queries concerning the owner.

2.2.2. Situational Context

In the restaurant domain, what the user looks for in a restaurant depends on why the user is eating out. If the user is alone, the preferences reflect the user's personal preferences. If the user is with their partner, the choice of restaurant reflects their joint preferences. When choosing a restaurant for a social gathering, the user may use different criteria in selecting a restaurant. While it is possible that these preferences would all be the same, this is unlikely to be true in general.

2.2.3. Location Context

Holidays are a good example of user preferences with location context. The user has a set of holiday preferences in their user profile. Some of these reflect the type of holiday the user prefers, e.g. skiing or beach. Some reflect where the user prefers to go, e.g. visiting historical sites from the Roman Empire. Thus, a location specified in the query would constrain which preferences would be applicable. For example, if the user were visiting Bali, then neither the Roman Empire nor the skiing preferences would apply, but the beach preference will.

2.2.4. Temporal Context

This should not be confused with the gradual evolution of user preferences over time. For example, the user is looking for a holiday in Switzerland. In summer, the user could prefer to go hill walking, while in winter, the user might be interested in skiing and other winter sports. Here, different user profile attributes may be selected depending on the time of year. A subtler form of temporal context is the concept of home. For most individuals, this is a fixed location. However, a student may live away from the parental home during term time and at the parental home at other times. Thus, the definition of home changes with the time of year.

2.2.5. Device Context

The user may have a set of display preferences. These may be generic, or domain dependent, or even situation dependent. There are two ways in which the display device influences the user preferences. Firstly, the user may have preferences that are specific to a particular device; e.g. the user may prefer a smaller font when using a mobile phone. Secondly, the device properties may prevent the use of the preferences, e.g. the user may state that they would like to see any video clips that are applicable but the device the user is currently using does not support this.

2.2.6. Multiple contexts

For some queries, more than one of the above contexts may be applicable. It is feasible for a user to be looking for a restaurant with their family while on holiday. Here, there are two situational contexts, namely on holiday and with family. Similarly, a student at home may be looking for a restaurant for their family. This query has a situational (family) and a temporal (is it term time?) context.
3. Using Context

When sharing user profiles, their separate contexts will almost inevitably lead to conflicts in user preferences; e.g. the restaurant preferences when dining on business compared to lunching with a young family. It is, therefore, important to capture and understand the context within the current query, within each user profile, and within the information sources before being able to produce a personalised response.

3.1. Conflicts

Conflicts in user profile data are a key issue in sharing profiles. Possible reasons for conflict may include:

- Two different user profiles with exactly the same data could result in different attribute values because of the perspective of the information provider. Thus two different user profiles could assign the same attribute a different priority. The problem is then to identify which is more important to the user in the current context.

- Information supplied to generate the profiles may come from different query contexts. For example, one profile may say that the user prefers a non-smoking restaurant while the other says that the user prefers one that permits smoking. However, closer examination may reveal that the first attribute has the context of being with their family, while the second attribute has the context of being alone. Thus, there is in fact no conflict between these two particular attributes. With knowledge of the context of profile attributes, it may be possible to resolve such conflicts.

- There may be inconsistencies in the data that the user supplies for the purposes of the modelling process due to errors on the part of the user.

   It is essential to be able to resolve conflicts quickly and reliably. To do so additional information is required, including knowledge about context. The aim of resolving conflicts is to deduce which (if any) of the attributes can be used for the current query context.

   The form a conflict takes may be a straight disagreement in the values of the attribute; the smoking/non-smoking example above shows how the user context suggests which value may be relevant to a given query. Another form of conflict is the priority of an attribute. As an example, one profile attribute may say that the user prefers information on restaurants that have non-smoking tables to be displayed first, while another states that the user will not consider any restaurant that does not offer a non-smoking table. Knowing the context in which the user supplied the information or in which the provider gathered the information could suggest a means of identifying the correct value to use for the current query.

   Although it may not always be possible to resolve a conflict, the use of context information will help in this process by reducing the number of options before seeking clarification from the user.

3.2. Using Context Information to Infer User Preferences

Consider the example of a restaurant profile generated in a restaurant in California. All restaurants in California are non-smoking, due to local laws. As such, a user profile generated in California is unlikely to have information about smoking preferences. However, it would be possible to infer that the user goes to non-smoking restaurants due to the location context information. This information
could then be used to seek a clarification for using this piece information out of context.

In Section 2.2, reference was made to profile attribute preferences that are absolute. An absolute preference can be used across several domains and also all contexts inside these domains. An example may be that of a physical disability of the user. Apart from absolute preferences, a preference is either relative or not applicable. A relative preference, e.g. a non-smoking attribute, is not always applicable in the domains different from the one in which they were generated, not in different contexts. However, this may vary between individual users, information that is relative for one may be absolute for others. If it can be deduced that a profile attribute is absolute (e.g. the user will not accept a smoking option if a non-smoking one is not available), then attributes of this type can be used in the same way as other absolutes. The problem with this is that an attribute may be absolute in some domains but not others. The ability to be able to deduce this is a challenge when inferring user profile information.

4. Conclusion

Conflict resolution (section 4.1) is one use that is envisioned for context information in the prototype being developed with this research. The first stage involves the use of a simple matching solution. If there is conflict, those items that do not match the current context are eliminated. This is the current stage of work. Future work involves looking at how context information can be used to infer preferences to resolve a conflict (section 4.2).

The assumption behind this work is that in the future the sharing of user profiles will be a solution to future personalisation problems. In order to consider profile sharing, a number of problems need to be overcome. Capturing context information is one such problem. This paper includes a first attempt at a classification of context, linking this to the problem of conflicts which arise when attempting to share profiles. The ability to identify conflicting knowledge and then resolve it is as important as the ability to use the knowledge correctly.

The ability to provide the user with greater levels of personalisation is a major challenge. Given that an increasing number of applications and information providers create user profiles, the solution put forward in this paper is to share these separate user profiles. Introducing context information, it becomes possible to select only the profile attributes that are useful for the current query.

Acknowledgements

The authors wish to acknowledge the support given to the work by the EPSRC through the DIP grant and the studentship for C. King.

References


Richard P. Smiraglia  
Palmer School of Library and Information Science, Long Island University, Brookville, NY, USA

Crossing Cultural Boundaries: Perspectives on the Popularity of Works

Abstract: Works are key entities in the universe of recorded knowledge. Works are those deliberate creations (known variously as opera, œuvres, Werke, etc.) that stand as the formal records of knowledge. Core bodies of works—canons—function to preserve and disseminate the parameters of a culture. There is some evidence that popularity of works is a contributing factor to the phenomena of mutation and derivation. In particular, novels demonstrated the high incidence of both derivation and mutation of a popular literary work. Commercial interests combine with cultural forces to propel the evolution of popular novels as interest in them spreads across language and geographic boundaries. Earlier studies support a concept of the work as a collaborative entity that is changed over time by those who embrace it. The more popular the work, the more likely we will observe change over time. Cultural boundaries are crossed by the mutation of best-selling works, as their translations find collaborative roles in cultures different from that in which the work originated. The study of works that have appeared on best-seller lists (one measure of cultural embrace, or "popularity") might yield useful data for comprehension of the content and extent of the canon of popular works. In the present study, a sample of best-selling works (fiction and non-fiction) from 1900-1999 is examined. Preliminary evidence from the first phase of this study demonstrates the consistency of the theoretical functioning of works as cultural entities. That is, works that enter a canon derive and mutate dramatically, while those that do not remain unchanged. "Popularity" is not demonstrated as equated with "best-selling," but all best-selling books in the sample generated more than one edition.

1. Introduction: Works as Entities of Recorded Knowledge

Works are key entities in the universe of recorded knowledge. Works are those deliberate creations (known variously as opera, œuvres, Werke, etc.) that stand as the formal records of knowledge. Core bodies of works—canons—function to preserve and disseminate the parameters of a culture. In the information retrieval domain, the work (as opposed to the document), has only recently received focused attention. Yet explicit linkage of relationships among entities is critical for information retrieval. Therefore, research to document the parameters of and relationships among works is critical for the advancement of knowledge organization.

Works, defined as entities for information retrieval, are seen to constitute sets of varying instantiations of abstract creations. Earlier studies (summarized in Smiraglia 2001) quantified the extent of instantiations of works, demonstrating patterns of mutation and derivation. There is some evidence that popularity of works is a contributing factor to the phenomena of mutation and derivation. In particular, novels demonstrated the high incidence of both derivation and mutation of a popular literary work. Commercial interests combine with cultural forces to
propel the evolution of popular novels as interest in them spreads across language and geographic boundaries.

Semiotic analysis also demonstrates the cultural role of works, which is further demonstrated by the network of relationships among instantiations (Leazer and Fumer (1999) term this a "textual identity network"). These variant instantiations and the networks among them must be explicitly identified in future systems for documentary information retrieval. An expanded perception of works helps us understand the variety of ways in which mechanisms for their control and retrieval might be shaped in future.

Earlier studies support a concept of the work as a collaborative entity that is changed over time by those who embrace it. One way to think of this phenomenon is this: the more popular the work, the more likely we will observe change over time. The more we can observe cultural embrace of a work, the more likely we can also observe change in the texts and instantiations of the work over time. Qualitative analysis by Leazer and Smiraglia (1999) suggested the same effect could be observed among sets of instantiations of works (bibliographic families, or textual identity networks) that are truly "popular" in the colloquial sense. Cultural boundaries are crossed by the mutation of best-selling works, as their translations find collaborative roles in cultures different from that in which the work originated. Derivations, such as screenplays and motion picture productions, demonstrate the cultural embrace of works; as their instantiations cross cultural and media boundaries. Integration of knowledge representation and organization demands better comprehension of the dimensions of works, of their patterns of derivation and mutation, and increased development of taxonomies of their types.

Consequently, it seemed appropriate to study works that have appeared on best-seller lists. One measure of cultural embrace, or "popularity," might reasonably be considered to be the judgment of the marketplace by which works wind up on best-seller lists. Analysis of best-sellers could potentially yield useful data for comprehension of the content and extent of the "canon" (as it were) of popular works. Evidence of the derivation and mutation of best-sellers can also augment our understanding of the social role of works as cultural signs. Thus, the over-riding research question for the study reported here was: Will best-selling books demonstrate patterns of derivation and mutation similar to those observed among works in other literary or academic "canons"?

2. Methodology

In the present study, a sample of best-selling works (fiction and non-fiction) from 1900-1999 is examined. For each work selected into the sample, all instantiations that have been described by library catalogs were identified using the online bibliographic utilities of OCLC and RLIN. Additionally, an attempt was made to identify any instantiations appearing on the world-wide-web. The resulting bibliographic families (or textual identity networks) can be analyzed to suggest the extent of derivation and mutation that characterizes the cultural life-cycle of best-selling works. Specifically, a frame of best-selling works was assembled by downloading titles for the ten best-selling books from 1900-1995: these lists are assembled from Publisher's Weekly's list of best-selling hardcover books for the entire century and are available at http://www.caderbooks.com/bestintro.html.
Fiction and non-fiction initially were placed in separate lists, though these were combined into a single list for the purposes of stratified random sampling.

There are no data from the earlier studies on the proportion of works that are best-selling books that have demonstrated patterns of derivation. Therefore, it was decided to use an approximation for the first phase of the study, from which better estimates could be generated for a more extensive second-phase analysis. One useful approximation to the proportion of derivation of best-sellers is the relative proportions of derivative works (in earlier studies) that are fiction and non-fiction. In Smiraglia (1992) 8.3% of the works were fiction demonstrating change over instantiations over time; 5.1% were narrative non-fiction. In the study reported by Smiraglia and Leazer (1999) 7.7% were fiction and 1.3% were narrative non-fiction. Sample size calculations suggested that 105 cases would be required for analysis of fiction alone, but only 51 would be required for analysis of nonfiction. Another useful approximation for the first phase of study is the reported size of bibliographic families. In Smiraglia (1992) the mean was 4.7; sample-size calculation using this estimate suggested a sample of 84 cases would be sufficient to yield useful results. Thus, a sample of 84 cases was drawn at random from the entire list of best-sellers. The sample was subsequently stratified into Fiction (45 cases) and Nonfiction (39 cases).

For each work in the sample, all bibliographic records were identified in both OCLC and RLIN. An attempt was also made to locate instantiations on the world-wide-web. The collected bibliographic records were then examined and codified (as in the earlier studies) into derivative relationship-types (see Appendix 3 in Smiraglia 2001 for details about this methodology). SPSS is used to provide statistical data for quantitative analysis. Phase II of the study will extend the sample based on better estimates of the rate of derivation. Phase III of the study will employ qualitative analysis of selected cases, and content analysis of the titles of the entire frame.

3. Characteristics of Best-Selling Books 1900-1995

A simple read through the lists of best-selling books from the twentieth century provides a qualitative summary of social and cultural trends throughout the century. The lists begin with romantic and cowboy fiction, and proceed through tales of rural and urban life and love. The nonfiction list moves from themes of women's suffrage through the whimsy of the late 1920s and early 1930s, but includes such icons of culture as General Marshall's report on post-World War II Europe, the Kiney's report on sexual function, and Woodward and Bernstein's account of the Watergate affair (not to mention quite a few cookbooks). The lists of works in the sample for phase 1 appear in tables 1 and 2 below.
<table>
<thead>
<tr>
<th>Year</th>
<th>Rank</th>
<th>Title</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>1901</td>
<td>3</td>
<td>The Helmet of Navarre</td>
<td>Bertha Runkle</td>
</tr>
<tr>
<td>1902</td>
<td>1</td>
<td>The Virginians</td>
<td>Owen Wister</td>
</tr>
<tr>
<td>1907</td>
<td>3</td>
<td>The Port of Missing Men</td>
<td>Meredith Nicholson</td>
</tr>
<tr>
<td>1907</td>
<td>10</td>
<td>Half a Rouge</td>
<td>Harold MacGrath</td>
</tr>
<tr>
<td>1912</td>
<td>7</td>
<td>The Just and the Unjust</td>
<td>Vaughan Kester</td>
</tr>
<tr>
<td>1913</td>
<td>9</td>
<td>The Valiants of Virginia</td>
<td>Hallie Erminie Rive</td>
</tr>
<tr>
<td>1914</td>
<td>7</td>
<td>Penrod</td>
<td>Booth Tarkington</td>
</tr>
<tr>
<td>1917</td>
<td>2</td>
<td>The Light in the Clearing</td>
<td>Irving Bacheller</td>
</tr>
<tr>
<td>1917</td>
<td>4</td>
<td>The Road to Understanding</td>
<td>Eleanor H. Potter</td>
</tr>
<tr>
<td>1917</td>
<td>5</td>
<td>Wildfire</td>
<td>Zane Grey</td>
</tr>
<tr>
<td>1917</td>
<td>6</td>
<td>Christine</td>
<td>Alice Cholmondeley</td>
</tr>
<tr>
<td>1917</td>
<td>9</td>
<td>The Definite Object</td>
<td>Jeffrey Farnol</td>
</tr>
<tr>
<td>1918</td>
<td>1</td>
<td>The UP Trail</td>
<td>Zane Grey</td>
</tr>
<tr>
<td>1918</td>
<td>2</td>
<td>The Tree of Heaven</td>
<td>May Sinclair</td>
</tr>
<tr>
<td>1920</td>
<td>10</td>
<td>Harriet and the Piper</td>
<td>Kathleen Norris</td>
</tr>
<tr>
<td>1923</td>
<td>1</td>
<td>Black Oxen</td>
<td>Gertrude Atherton</td>
</tr>
<tr>
<td>1925</td>
<td>1</td>
<td>Soundings</td>
<td>A. Hamilton Gibbs</td>
</tr>
<tr>
<td>1927</td>
<td>5</td>
<td>Jalna</td>
<td>Mazo de la Roche</td>
</tr>
<tr>
<td>1928</td>
<td>7</td>
<td>Old Pybus</td>
<td>Warwick Deeping</td>
</tr>
<tr>
<td>1933</td>
<td>3</td>
<td>Ann Vickers</td>
<td>Sinclair Lewis</td>
</tr>
<tr>
<td>1933</td>
<td>6</td>
<td>Porgive us our Trespasses</td>
<td>Lloyd C. Douglas</td>
</tr>
<tr>
<td>1933</td>
<td>7</td>
<td>The Master of Jalna</td>
<td>Mazo de la Roche</td>
</tr>
<tr>
<td>1936</td>
<td>10</td>
<td>Eyeless in Gaza</td>
<td>Aldous Huxley</td>
</tr>
<tr>
<td>1939</td>
<td>1</td>
<td>The Grapes of the Wrath</td>
<td>John Steinbeck</td>
</tr>
<tr>
<td>1941</td>
<td>4</td>
<td>The Sun is my Undoing</td>
<td>Marguerite Steen</td>
</tr>
<tr>
<td>1942</td>
<td>5</td>
<td>Drivin’ Woman</td>
<td>Elizabeth Pickett</td>
</tr>
<tr>
<td>1946</td>
<td>10</td>
<td>The Snake Pit</td>
<td>Mary Jane Ward</td>
</tr>
<tr>
<td>1949</td>
<td>9</td>
<td>Pride’s Castle</td>
<td>Frank Yerby</td>
</tr>
<tr>
<td>1958</td>
<td>10</td>
<td>Victorine</td>
<td>Frances Parkinson Keyes</td>
</tr>
<tr>
<td>1960</td>
<td>6</td>
<td>The Constant Image</td>
<td>Marcia Davenport</td>
</tr>
<tr>
<td>1965</td>
<td>6</td>
<td>Those who love</td>
<td>Irving Stone</td>
</tr>
<tr>
<td>1966</td>
<td>10</td>
<td>All in the Family</td>
<td>Edwin O’Connor</td>
</tr>
<tr>
<td>1969</td>
<td>2</td>
<td>The Godfather</td>
<td>Mario Puzo</td>
</tr>
<tr>
<td>1970</td>
<td>9</td>
<td>Travels with my Aunt</td>
<td>Graham Greene</td>
</tr>
<tr>
<td>1978</td>
<td>1</td>
<td>Chesapeake</td>
<td>James A. Michener</td>
</tr>
<tr>
<td>1979</td>
<td>5</td>
<td>Jailbird</td>
<td>Kurt Vonnegut</td>
</tr>
<tr>
<td>1981</td>
<td>5</td>
<td>Gorky Park</td>
<td>Martin Cruz Smith</td>
</tr>
<tr>
<td>1985</td>
<td>8</td>
<td>Lucky</td>
<td>Jackie Collins</td>
</tr>
<tr>
<td>1986</td>
<td>1</td>
<td>It</td>
<td>Stephen King</td>
</tr>
<tr>
<td>1987</td>
<td>7</td>
<td>Presumed Innocent</td>
<td>Scott Turow</td>
</tr>
<tr>
<td>1988</td>
<td>8</td>
<td>To be the best</td>
<td>Barbara Taylor Bradford</td>
</tr>
<tr>
<td>1989</td>
<td>4</td>
<td>Star</td>
<td>Danielle Steel</td>
</tr>
<tr>
<td>1994</td>
<td>2</td>
<td>Debt of Honor</td>
<td>Barbara Taylor Bradford</td>
</tr>
<tr>
<td>1996</td>
<td>3</td>
<td>Desperation</td>
<td>Stephen King</td>
</tr>
<tr>
<td>1996</td>
<td>4</td>
<td>Airframe</td>
<td>Michel Crichton</td>
</tr>
</tbody>
</table>

Table 1. Fiction Best-Sellers in Phase I Sample
<table>
<thead>
<tr>
<th>Year</th>
<th>Rank</th>
<th>Title</th>
<th>Author</th>
<th>Year</th>
<th>Rank</th>
<th>Title</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>1901</td>
<td>3</td>
<td>The Helmet of Navarre</td>
<td>Bertha Runkle</td>
<td>1912</td>
<td>9</td>
<td>Woman and Labor</td>
<td>Olive Schreiner</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>White Shadows in the South Seas.</td>
<td>Frederick O'Brien</td>
</tr>
<tr>
<td>1902</td>
<td>1</td>
<td>The Virginian</td>
<td>Owen Wister</td>
<td>1920</td>
<td>5</td>
<td>The Port of Navarre</td>
<td>Margot Asquith</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Labor. Schreiner</td>
<td></td>
</tr>
<tr>
<td>1907</td>
<td>3</td>
<td>The Port of Missing Men</td>
<td>Meredith Nicholson</td>
<td>1921</td>
<td>4</td>
<td>The Autobiography of Margot Asquith.</td>
<td>Emile Coué</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1907</td>
<td>10</td>
<td>Half a Rogue</td>
<td>Harold MacGrath</td>
<td>1923</td>
<td>7</td>
<td>Self-Mastery Through Conscious Auto-Suggestion.</td>
<td>Bernard Shaw</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>George Hackett</td>
</tr>
<tr>
<td>1912</td>
<td>7</td>
<td>The Just and the Unjust</td>
<td>Vaughan Kester</td>
<td>1924</td>
<td>8</td>
<td>Saint Joan.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bernard Shaw</td>
</tr>
<tr>
<td>1913</td>
<td>9</td>
<td>The Valiants of Virginia,</td>
<td>Hallie Rives</td>
<td>1928</td>
<td>10</td>
<td>The Intelligent Woman's Guide</td>
<td>Bernard Shaw</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>to Socialism and Capitalism.</td>
<td></td>
</tr>
<tr>
<td>1914</td>
<td>7</td>
<td>Penrod</td>
<td>Booth Tarkington</td>
<td>1931</td>
<td>7</td>
<td>Fatal Interview.</td>
<td>Edna St. Vincent</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Millay</td>
</tr>
<tr>
<td>1917</td>
<td>2</td>
<td>The Light in the Clearing</td>
<td>Irving Bachelier</td>
<td>1932</td>
<td>3</td>
<td>A Fortune to Share.</td>
<td>Vash Young</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1917</td>
<td>4</td>
<td>The Road to Understanding</td>
<td>Eleanor Porter</td>
<td>1935</td>
<td>3</td>
<td>Life with Father.</td>
<td>Francis Hackett</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Francis the First.</td>
<td></td>
</tr>
<tr>
<td>1917</td>
<td>5</td>
<td>Wildfire</td>
<td>Zane Grey</td>
<td>1935</td>
<td>6</td>
<td>Man the Unknown.</td>
<td>Alexis Carrel</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>New England: Indian Summer.</td>
<td></td>
</tr>
<tr>
<td>1917</td>
<td>6</td>
<td>Christine</td>
<td>Alice Cholmondeley</td>
<td>1936</td>
<td>1</td>
<td>The White Cliffs.</td>
<td>Alice Duer Miller</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ten Years in Japan.</td>
<td></td>
</tr>
<tr>
<td>1917</td>
<td>9</td>
<td>The Definite Object</td>
<td>Jeffrey Farnol</td>
<td>1940</td>
<td>7</td>
<td>General Marshall's Report.</td>
<td>Joseph C. Grew</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1918</td>
<td>1</td>
<td>The UPTail</td>
<td>Zane Grey</td>
<td>1941</td>
<td>2</td>
<td>The Fireside Book of Folk Songs.</td>
<td>Margaret B. Boni</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1918</td>
<td>2</td>
<td>The Tree of Heaven Harriet and the Piper</td>
<td>May Sinclair, Kathleen Norris</td>
<td>1944</td>
<td>10</td>
<td>Sexual Behavior in the Human Male</td>
<td>A.C. Kinsey et al.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1920</td>
<td>10</td>
<td>Jaha</td>
<td>Mazo de la Roche</td>
<td>1950</td>
<td>3</td>
<td>Look Younger, Live Longer.</td>
<td>Gayelord Hauser</td>
</tr>
<tr>
<td>Year</td>
<td>Page</td>
<td>Title</td>
<td>Author(s)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>---------------------------------</td>
<td>----------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1928</td>
<td>7</td>
<td>Old Pybus</td>
<td>Warwick Deeping</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1933</td>
<td>3</td>
<td>Ann Vickers,</td>
<td>Sinclair Lewis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1933</td>
<td>6</td>
<td>Forgive Us Our Trespasses,</td>
<td>Lloyd C. Douglas</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1933</td>
<td>7</td>
<td>The Master of Jaina,</td>
<td>Mazo de la Roche</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1936</td>
<td>10</td>
<td>Eyeless in Gaza,</td>
<td>Aldous Huxley</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1939</td>
<td>1</td>
<td>The Grapes of Wrath,</td>
<td>John Steinbeck</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1941</td>
<td>4</td>
<td>The Sun Is My Undoing,</td>
<td>Marguerite Steen</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1942</td>
<td>5</td>
<td>Drivin' Woman,</td>
<td>Elizabeth Pickett</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1946</td>
<td>10</td>
<td>The Snake Pit, Mary Jane Ward</td>
<td>Mary Jane Ward</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1949</td>
<td>9</td>
<td>Pride's Castle, Frank Yerby</td>
<td>Frank Yerby</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1958</td>
<td>10</td>
<td>Victorine,</td>
<td>Frances Parkinson Keyes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1960</td>
<td>6</td>
<td>The Constant Image</td>
<td>Marcia Davenport</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1965</td>
<td>6</td>
<td>Those Who Love,</td>
<td>Irving Stone</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1966</td>
<td>10</td>
<td>All in the Edwin Family,</td>
<td>O'Connor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1969</td>
<td>2</td>
<td>The Godfather,</td>
<td>Mario Puzo</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1970</td>
<td>9</td>
<td>Travels with My Aunt,</td>
<td>Graham Greene</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1978</td>
<td>1</td>
<td>Chesapeake,</td>
<td>James A. Michener</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1979</td>
<td>5</td>
<td>Jailbird,</td>
<td>Kurt Vonnegut</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1981</td>
<td>5</td>
<td>Gorky Park,</td>
<td>Martin Cruz Smith</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1950</td>
<td>9</td>
<td>Campus Zoo.</td>
<td>Clare Barnes Jr.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1955</td>
<td>5</td>
<td>How to Live 365 Days a Year.</td>
<td>John A. Schindler</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1958</td>
<td>9</td>
<td>Dear Abby.</td>
<td>Abigail Van Buren</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1959</td>
<td>7</td>
<td>The Elements of Style.</td>
<td>William Strunk Jr., E. B. White</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1960</td>
<td>5</td>
<td>Better Homes and Gardens Dessert Book</td>
<td>compiled by Bill Adler Rod McKuen David Reuben, M.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1964</td>
<td>7</td>
<td>The Kennedy Wit.</td>
<td>compiled by Bill Adler Rod McKuen David Reuben, M.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1969</td>
<td>10</td>
<td>Twelve Years of Christmas.</td>
<td>Any Woman Carl!</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1971</td>
<td>5</td>
<td>The President's Men.</td>
<td>Alistair Cooke's America.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1973</td>
<td>9</td>
<td>Alistair Cooke's America.</td>
<td>All the President's Men.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1976</td>
<td>1</td>
<td>The Final Days.</td>
<td>The Final Days. Bob Woodward and Carl Bernstein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1979</td>
<td>7</td>
<td>Lauren Bacall by Myself.</td>
<td>Lauren Bacall by Myself.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1983</td>
<td>4</td>
<td>The One Minute Manager.</td>
<td>The One Minute Manager.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1994</td>
<td>1</td>
<td>In the Kitchen with Rosie.</td>
<td>In the Kitchen with Rosie.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1995</td>
<td>7</td>
<td>Mars and Venus in the</td>
<td>Mars and Venus in the</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Non-fiction Best-Sellers in Phase I Sample

<table>
<thead>
<tr>
<th>Year</th>
<th>Rank</th>
<th>Title</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>1985</td>
<td>8</td>
<td>Lucky,</td>
<td>Jackie Collins</td>
</tr>
<tr>
<td>1986</td>
<td>1</td>
<td>It,</td>
<td>Stephen King</td>
</tr>
<tr>
<td>1987</td>
<td>7</td>
<td>Presumed Innocent,</td>
<td>Scott Turow</td>
</tr>
<tr>
<td>1988</td>
<td>8</td>
<td>To Be the Best,</td>
<td>Barbara Taylor</td>
</tr>
<tr>
<td>1989</td>
<td>4</td>
<td>Star,</td>
<td>Danielle Steel</td>
</tr>
<tr>
<td>1994</td>
<td>2</td>
<td>Debt of Honor,</td>
<td>Tom Clancy</td>
</tr>
<tr>
<td>1996</td>
<td>3</td>
<td>Desperation,</td>
<td>Stephen King</td>
</tr>
<tr>
<td>1996</td>
<td>4</td>
<td>Airframe,</td>
<td>Michael Crichton</td>
</tr>
</tbody>
</table>

The sampling frame contained 1836 works (992 fiction and 844 nonfiction). Some authors repeated (Winston Churchill, Danielle Steele, James Michener, for example), but not many. 178 authors of fiction appear on the list more than once, and 62 works repeat as well (one four times) indicating a small number of works that remain best-sellers over time. At least three works can be identified as having spawned sequels that also appeared on the list. Cultural influences have impact on the list as well; for example, during World War I separate lists of "war books" were ranked alongside the regular best-sellers.

4. Preliminary Results

At the time of this writing data analysis is in too preliminary a stage to allow reporting of specific results. Nonetheless, certain parameters of the data are already apparent. In earlier studies of works (see Smiraglia 2001), work were seen to have the same set of characteristics:

- The majority exist in only one edition, but substantial proportions generate bibliographic families through mutation and derivation over instantiations over time;
- Simultaneous and successive editions, and translations, predominate among the types of mutation and derivation observed;
- Older progenitors are associated with the largest bibliographic families;
- More recent progenitors are associated with the most complex bibliographic families.

Preliminary analysis indicates confirmation of these trends in the present study as well. One major exception: every work but one in the sample demonstrated mutation and derivation over instantiations over time. That is, only one work existed in only its initial edition. Best-selling works, therefore, demonstrate wider cultural impact through the promulgation of editions and translations over time than is true of the works observed in earlier studies of academic canons. However, the majority of best-selling works in the sample exist in a small set of instantiations: the progenitor, one or two simultaneous editions, and one or two successive editions.
That is, even best-selling works demonstrate a concept of "canonicity," only some works are culturally embraced sufficiently to generate large bibliographic families. Web-searching was employed for the first time in the present study to begin to understand how the multiple instantiation of works might or might not present challenges in the digital environment. Preliminary data are inconclusive: few works in the sample are represented in full in the digital environment.

5. Epistemological Implications

Epistemology for documentary analysis provides key perceptual information about the objects of knowledge organization. Works are carriers of knowledge, representing deliberately-constructed packages of both rational and empirical evidence of human knowledge. Smiraglia (2002) explicates the epistemological parameters of works viewed as entities for information retrieval. A taxonomy of mutations and derivations yielded from empirical observation and qualitative analysis leads to epistemological understanding of the cultural roles of works. In earlier studies qualitative analysis of complex bibliographic families revealed the ways in which mutation and derivation types interact. Analysis of the present study is in an early phase, but one example can be outlined. Owen Wister's *The Virginian* was published as a novel in 1902 in New York. Over time, this work has been observed in 134 instantiations, including 89 successive editions, 14 translations, at least 20 performances, and a variety of adaptations, including the generation of a comic-book version based on a set of excerpts. Many of these instantiations demonstrate more than one type of derivation. For instance, the successive editions often are issued simultaneously (new editions might be issued in three cities at once each year). These patterns demonstrate both the working of the marketplace and the cultural embrace of the work. Further analysis will demonstrate whether a model for integrated representation of works and organization of their representations for retrieval can emerge.

6. Theoretical Implications

The present research seems to promise to add to the growing body of empirical evidence about the derivation and mutation of works in instantiations over time. Repeated observation of quantitatively demonstrable phenomena supports the notion that a theory of the instantiation of works over time can be developed. Such a theory would have predictive power critical for the design of sophisticated work-retrieval engines. Qualitative analysis also promises to contribute to understanding of works and their social roles. Eggert (1994) suggested that works the gain wide cultural acceptance become collaborative ventures; this effect seems to be demonstrated by the forward spiral of mutating and deriving instantiations over time. Such approaches to grounded theory can also prove critical for the design of sophisticated work-retrieval engines.

References


A. Neelameghan and Hemalata Iyer
School of Information Science and Policy, University at Albany, SUNY, Albany, USA

Some Patterns of Information Presentation, Organization, and Indexing for Communication Across Cultures and Faiths

Abstract: Converging digital technologies and networks, especially the Internet offer considerable scope for individuals and organizations to be authors and publishers with the possibility of getting global response to one's ideas. But there are also several impediments to overcome - existence of a multiplicity of languages, cultural bias, misinterpretation of concepts, non-existence and/or non-acceptance of ideas of one group by people of other cultures and faiths. This paper examines the types of indexes, glossary of terms, organization and presentation of text in different languages/scripts, transliteration of selected text, commentaries, help to correct pronunciation, listing of common errors in rendering text, song, etc. by beginners and people of different cultures and language groups, etc. as means and approaches to mitigate the difficulties of communication across cultures, faiths, and linguistic boundaries. Such provision in typical examples of information resources in the spiritual field - an Internet site, and two CD multimedia - is discussed.

1. Impact of Information Technologies

The converging digital technologies make it possible for almost seamless access to vast volumes of information, in a variety of media, on a global scale. The Internet offers the potential for almost anyone to be an author and publisher and to get one's ideas disseminated nationally and internationally, with the possibility of global response. Hence, organizations of all types and even individuals are taking advantage of this potential. On the positive side, it facilitates cross-cultural communication fostering better understanding among peoples across national, linguistic, and cultural boundaries. Fostering such exchange of ideas is important for the development of all fields of human endeavour and, perhaps, most crucial, in the religious, spiritual and cultural fields. Since conflicts arise in the minds of people, due to misinterpretation and misunderstanding of ideas and practices of one's own and more importantly other people's philosophies, religious tenets, beliefs and cultures, advantage can be taken of global communication facilities offered by digital technologies. There are several efforts in this regard. But there are also several apparent impediments to overcome - existence of a multiplicity of languages, cultural bias, misinterpretation of concepts, non-existence and/or non-acceptance of ideas and practices of one culture or faith in other cultures and faiths, and so on. Provision of different types of indexes, glossary of terms, organization and presentation of text in different languages/scripts, transliteration of select text, commentaries, help to correct pronunciation, listing out common errors in rendering text, song, etc. by beginners and people of different cultures, language groups, links to other websites and literature sources are among the strategies and approaches to mitigate the problems of communication across cultures, faiths and linguistic boundaries.
2. Scope of the Paper
This paper discusses the devices mentioned in section 1 in typical examples of information resources in the spiritual field - an Internet site, and two CD-ROM multimedia sources1. The provisions and devices in these sources are intended to contribute to a correct and contextual interpretation of ideas and practices across linguistic and cultural boundaries.

3. Examples
The contents of this Internet site (maintained by Debb Platt) relate essentially to mysticism (the nature of reality, the individual's struggle to attain a clear vision of reality, and the accompanying transformation of consciousness. The “site explores the mystical traditions of six religions” (Judaism, Christianity, Islam/Sufism, Hinduism, Buddhism, Taoism) by comparing and contrasting quotations drawn from their respective literature.

The site consists of hundreds of linked web pages. Starting with a specific concept in one religion for which definition and comments are provided, one can click on specialized terms for further information from another religion to read about equivalent and related concepts in that religion. Definitions are provided for the basic concepts in a hyperlinked Glossary database. One can focus either on one religion or one mystic and then navigate across using the hypertext links. Links to selected source materials are provided, facilitating, among other things, comparative study of concepts of mysticism in the six religions.

The quotations are organized by topic. The group headings and subheadings of the topics and their arrangement appear to parallel the ‘individual’s approach to attain a clear vision of Reality and the stages of transformation accompanying that vision’.

- There's reality beyond the material world:
  - Which is uncreated
  - It pervades everything
  - but remains beyond the reach
  of human knowledge and understanding

- You approach that reality by:
  - Distinguishing ego from true self
  - Understanding the nature of desire
  - Becoming unattached
  - Forgetting about preferences
  - Not working for personal gain
  - Letting go of thoughts
  - Redirecting your attention
  - Being devoted
  - Being humble
  - Invoking that reality
  - Surrendering

- That reality approaches you through
  - Grace
  - The teacher

- You're transformed so that you embody that reality by:
  - Dying and being reborn
- Seeing the light
- Experiencing union
- Experiencing freedom

---

+Home Page +Mystics, theologians and scriptures +Glossary +Links to other sites
+What's new?

Each of the topic headings in italics can be clicked on to call the related extract for each of six religions. Examples are given below. Note the links to Glossary, Home Page, Other sites, etc.

Clicking on 'Which is uncreated' calls up

**Jewish: Ashlag: Uncreated**

Uncreated

(Quotations from Yehuda Ashlag)

Yehudas L. Ashlag:

... all names and appellations discussed in the wisdom of the Kabbalah concern only the Light extended from the Creator's substance and not the substance itself, for truly speaking, we have neither word nor thought of expression by which to speak of His substance.

(p. 56)

... with one "Thought" of the Creator all existence was emanated and created, the upperworlds in conjunction with the lower worlds, including the evolutionary processes of development which creation will constantly undergo until all functions reach their final completion - the Millennium...

(p. 37)

---

**Christian: Cyril: Uncreated**

Uncreated

(Quotations from Cyril of Alexandria)

The divine transcenence

... the difference between Creator and created is incomparable

(Cyril of Alexandria: Select Letters", p.103, quoting Cyril's "On the Creed" 10)

(God is) incorporeal, immaterial, impalpable, beyond quantity and circumscription, beyond form and figure

(The Image of God in Man according to Cyril of Alexandria, p.22, quoting "Responsiones ad Teberium" 14)

---

And more quotations for and from each of six religions. Some of the citations are hyper-linked to the source materials listings and/or brief biography of the author or details about cited source. Example:


The opening page of this multimedia CD (Swami Vivekananda, 2000) presents the following menu (one icon for each): *Benedicticon, Life, Research and Thinkers, Call to Humanity, Works, Gallery, For children, Centres, and About us.* The contents of each of these sections is detailed below.
The background Indian classical / devotional music creates a spiritual atmosphere. Generally, where applicable, in all screens, there are buttons for sound (voice or music), to link to glossary, for searching by index, for printing; to move up and down, next / previous page, contents, and display Sanskrit content with transliteration (in English). Text and pictures can be zoomed.

**Benediction:** About Swami Vivekananda, narration (voice) by Swami Ranganathananda (13th President of the Ramakrishna Mission), with text on screen; also pictures of Vivekananda and Ranganathananda.

**Life, Research and Thinkers:** Picture of shelved volumes - Life, 2v., Research, 6v., Thinkers, 1v. (See Works below).

**Works:** Pictures of nine shelved volumes. Click to open any of the volumes and in each case the contents page is displayed. A particular content heading may be clicked to open the corresponding page. Highlighted terms can be clicked to link to the Glossary for meaning.

**Call to Humanity:** Extracts from Vivekananda’s works and discourses organized under the following headings: Faith and Strength; Powers of the Mind; Man the Maker of his Destiny; Education and Society; Serve Man as God; Religion and Ethics; Glory of India; Other Exhortations. Each heading can be clicked to open corresponding quotation. Click icon to get citation of source.

**For Children:** Selected ‘stories’ from the Works suitable for children, arranged under 30 clickable headings.

**Quiz:** Test questions categorized as: True or false; Objective (questions); Match the Words; Match the Pictures.. Enrich your Knowledge-- Select a quiz category by clicking the heading. The last heading leads to an index to Vivekananda’s Life and Works.

**Gallery:** Some 275 pictures relating to Swami Vivekananda, and persons, places and other entities associated with him. Clicking on a particular picture displays an explanatory message with voice reading.

**About Us:** Brief notes about Advaita Ashramas located in different countries, with pictures, and online link to websites, if any.

**Centres:** World map with location of Ramakrishna Mission marked. Clicking on a marker displays address, note on activities, services offered, etc. at the Centre.

This CD targets children as well as scholars, laypersons interested in Advaita Philosophy. The thematic presentation, access to the contents of Vivekananda’s multi-volumed set, and the children’s stories along with a quiz are the significant features of this CD.

### 3.3 Saint Tyagaraja: A Multimedia Presentation (2001)

This multimedia CD (Saint Tyagaraja, 2001) presents a unique portrayal of 108 Tyagaraja devotional songs (kritis). The product is a joint effort of scholars of Carnatic music, musicians and information professionals. It attempts to especially foster understanding and appreciation of the kritis across linguistic borders within India. In addition, the presentation also promotes interest in Carnatic music for international audiences. The CD has several features that foster communication across linguistic boundaries in India and to some extent even internationally.

The 108 kritis included in the product represent different themes, ragas (melodic matrices or patterns), talas (rhythm or beats) and art models used by Tyagaraja. Thirty-five of the kritis are rendered in instruments - the veena, violin, and flute. It includes commentaries, meaning and text of the kritis that aid in understanding the meaning.
The opening screen has icons for:

**Help**
Searching database by title, theme, *ragam*,
*talam* (beat), and other fields

**Browse kritis database**
- Record by record
- See lyrical content
- See summary
- See commentary
- See text of *kritis*

**Index to kritis**
- See list of titles
- Sort index by Raga
- Sort index by *Tula*
- Select *kriti* to view
- Listen to music
- Listen to commentary

**Index to deities**
- See *kritis* on deities
- Select *kriti* to view

**Open Internet browser**
- Browse articles
- Link to websites
- Print text of *kritis*

**The music menu options include:**
- Select and play *kriti* (vocal)
- Select and play *kriti* (instrumental)
- Play title music
- Stop title music
- Continuous play
- Favourites

**Index to all lyrical content word**
Or index to common errors in articulation of content words
See meaning of words
Select *kriti* by content word

Slide show on quantitative analysis
See distribution of *kritis* by *ragas* and other tables

Slide show on Tyagaraja
- Early life
- Adult and middle life
- Works
- Tyagaraja literature
- Premonition and final end
- Tyagaraja Aradhana
- Festivals

**Database Searching:** The databases can be searched by title, theme, *ragam*, *talam* and other fields, such as, keyword, deity, vocalist and lyrical content. Every composition is provided with a special commentary, lyrical text, summary, notes, and the text in four languages / script: English / Roman (default), and the Indian languages, Telugu, Tamil, and Devanagiri. Further help is provided through a linked glossary accessible from the summary, lyrical content special commentary, and notes sections of the CD.

**The Lyrical Content** section shows a table of all Telugu lyrical content words/phrases and their meanings in English.

**The Summary** section provides a gist of the contents. Since many of Tygaraja’s compositions draw from the events in the Indian epic, *Ramayana*, and other Hindu scriptures, the less known stories alluded to, are presented in detail. This is particularly helpful to those unfamiliar with Indian culture, and also for Indians and people of Indian origin who may wish to gain more knowledge of the cultural heritage conveyed through the epics and scriptures.

**The Index to Common Errors in Rendering of Content Words:** This section presents an alphabetical listing of words / phrases from Telugu lyrics along with its meaning in English. It provides guidance on where the word should be split and how it should be pronounced while the song is musically rendered. A list of such common
errors along with how it aught to be sung is also demonstrated. This is of invaluable assistance to non-Telugu speaking singers and users. Besides the language barrier, some Carnatic music lovers may tend to pay more attention to the technicalities of music rather than to the lyrics.

**Commentary:** This section allows the users to view the text in four languages while music / commentary is played or heard.

**Special commentary** - presents information that fosters music appreciation, such as, the impact of raga, basic music technicalities, etc. Users can choose to play just the music, or music with commentary, or just plain commentary. For example, search by the theme ‘gratitude’ retrieves the kriti ‘Dasarathe ni momu.’ The special commentary provides insight into the deployment of the raga, the use of swaras and their effect, thereby promoting music appreciation. Though the technical terminology used in the commentary could prove as a barrier to lay-users in music, the glossary of terms can be accessed for meaning of these terms.

*The Index to Kritis* allows users to sort the database by the kriti title, raga and tala. Thus for instance, by selecting a particular raga, *Todi*, all the kritis set to *Todi* can be retrieved.

**Quantitative Analysis of Saint Tyagaraja’s Compositions:** A slide show presents a quantitative analysis of his 688 compositions by raga groups. This feature of the CD is especially of interest to Carnatic music scholars.

**Information on Tyagaraja:** The biographical sketch of the saint is provided by means of a Power-Point presentation titled *Saint Tyagaraja’s life and works*. In addition, the feature, *Tyagaraja on the Internet* provides links to specially commissioned articles, text of krithis, index to web sites and select bibliography on Tyagaraja. These are very educative to both national and international audiences.

### 4. Concluding Remarks

A variety of devices and methods of presentation of information - text, image, and sound (voice narration and singing) - are used to facilitate contextual interpretation of concepts, and cross-cultural communication. The presentation and devices used also vary with the intended target audience or users. The resources usually give explanation of specialized concepts and definition of terms in the text itself or with the aid of a hyperlinked glossary. Hypertext links and / or cross references to the sources of information used in preparing the databases are given. These also direct the user to other information sources.

**Notes:**

1. We were to include another multimedia CD (OM Information Service) and a printed source (*Srimad Bhagavatam, 18v*) as examples, but these have been omitted in this paper due to space restriction.

2. Example: When the phrase *Chinta thirchchuta enta modira* is selected, a brief commentary indicating that *Chintathirchutakenta* being a single word, should not be split as *Chintatirchchuta enta*. This is further illustrated by singing (musical rendition of) that line from the composition. This is of invaluable assistance to non-Telugu speaking singers and users. Besides the language barrier, some Carnatic music lovers may tend to pay more attention to the technicalities of music rather than to the lyrics.

3. The special commentary to the *kriti* “Dasarathe ni momu” reads as, “The kriti commences with adhara sadja, introduces both the depth in the raga and the impact of the lyrics. In combination with the predominant note, the panchamam, the fact is conveyed how, even without the use of the other dominant swaras, the raga todi can be portrayed effectively. The
two svaras sadja and panchamam play the major role. The svaras in the scale of the raga are all shuddha svaras exactly opposite to that of Kalyani. Even within the scale, one should note that the jivasvaras, gandhara has a slightly higher frequency in the arohana than in the avarohana. The song is set in the right pace, neither too slow nor too fast.”

References
Knowledge Representation from Amazonian Narratives: Culture and Oral Tradition

Abstract: Knowledge representation is related to several disciplines, as a field of interdisciplinary knowledge. The article presents a study including the text, the culture and the geographic space. The objective of the work is to show the representation of the knowledge studied in texts from popular verbal narratives to rescue culture with mapped terminologies in specific spaces of the Paraense Amazonia. The analysis is carried out both in the study of the term in narrative speeches as well as in the identification of the culture and the space where the history takes place. The classification of the codes of culture, its relation of the sign to the signs and the systems of signs is part of the collective cultural conscience to retrieve the information in the Age of knowledge to reach the universality.

1. Introduction

The integration of knowledge throughout borders is a concern in the current moment to assist better the needs of several societies. A great amount of information has been available through electronic media for the generation of new knowledge. The progress in text digitalization and the homepages are contributions for the society of knowledge. However, the great challenge is the retrieval of information via communication system, in order to generate knowledge. The retrieval of the information is one of the steps to a reflection on ways of how to represent the knowledge and organize the information to reach the universality. Many questions have appeared for a reflection and studies due to the change of society and of the concept of representation and organization of knowledge, in relation to the universality. How can that universality be represented in conceptual structures and integrated with specialized knowledge of cultures and geographical areas? This and other subjects are concerns that need to be studied in order to assist the several societies in the world.

In this work the model and the concept of knowledge representation come from a qualitative analysis in Amazonian narratives, showing the representation of the knowledge as an interdisciplinary field and as a communicative system in collective memory, starting from what people tell in different areas and cultures, having the aim of rescuing the popular culture, with mapped terminologies in the specific spaces of the Paraense Amazonia.

2. Knowledge Representation

The representation of the knowledge has been studied, for many specialists in the subject, to assist universality. It is an interdisciplinary field as well as a communicative system in collective memory. As a field of interdisciplinary knowledge, it has several conceptions and is linked to several disciplines, defined in function of their use. Oliveira (2001a) observes that the term representation of knowledge is present in different contexts, with several significances, implying
several forms of knowledge and representation. The knowledge comes from the information in order to be represented. The information is documented, both in the written text and in the subject's memory which is part of a certain world. The text, through the signs, transmits the thought of the people that writes or lives together in a certain space-time.

In that way, human beings produce knowledge that needs to be represented, to be retrieved and transferred to the society. However, as Morin (1996) observes, "the human cognitive system produces knowledge by building from the treatment of signals/signs/symbols, the translations which are the representations, speeches, ideas, theories". In that sense, knowledge is understood as action, built to be activity and product of its cognition, represented by a language of a certain culture. Man uses signs that represent his thoughts to communicate and to elaborate his culture. Therefore, the representation of the knowledge is a complex subject that demands a reflection in relation to the text type, the culture and the space where that culture develops, to get at the communicative system as collective memory.

In the communicative system, it is observed that in each space of human beings, a regime of signs is developed, a specific semiotic. Lévy (1998) shows that the sign is part of the being, and the being of the sign as well, in the world in which he lives. The speeches become acts, owing powers and man goes on developing the universe of significance. This is how the subject of knowledge learns and transmits from generation to generation, keeping the duration of knowledge. However, when knowledge is distributed in the collective subject (the net of the nets) it can be taken even further, independent of the territorial knowledge that is a reserved domain that leaves the peasants and the illiterate ones out (Lévy, 1998).

In the cultural policy, the territory is like that, one of the determinants of the cultural identity, beside of the constitution and preservation of collections. The territory is, let's say, the fact of having the individual who was born in an area still living in it. The individual establishes coexistence, by means of linguistic elaboration, the daily behavior and the cultural works, forming a symbolic construction, that is sometimes identified with certain elements, other times with different ones (Coelho, 1999, p. 201).

3. Narrative, Culture, Oral Tradition

The narrative is a text type in which the knowledge is studied to be represented. On it the culture and the oral tradition can be observed. This is because the oral narratives of people or social segments of traditional cultures, either from tribal or popular tradition, such as the myths, the legends, the tales, life histories, according to Borges (2001), can be considered genders of the verbal art. They are part of societies, both in the aesthetic and cultural formational of the peoples.

On this way, the verbal art will take us to a discursive chain, giving sense to the histories by means of the social imaginary and acting as element of identity connection, among various generations, as imaginary memory (Borges, 2001). Thus, the oral narratives, in the report conditions are good to understand the constitution of human societies, in its discursive materiality, the territory of the relationship, important for human beings and their temporalities. The populations of traditional culture can be analyzed as signs of the historical-ideological formation, noticed in the linguistics of the statement itself. Besides, we can notice that the
narratives repair the past where the notion of nature is not in opposition to the notion of culture. The cultural identity will point to:

"a representation system (symbolic elements and procedures of the acting of those elements) of the relationships between the individuals and the groups and between these and their reproduction and production territory, their habitat, their space and their time " (Coelho, 1999, p. 201).

In relation to culture it can be observed that it possesses several significances. Ferreira (1994/959) understands culture as memory. This author in her studies about Lotmam, as the thinker of the Târtu School and pioneer in the study in the field of the typology of culture, observes that her thesis as regards to typology of culture is that:

"it is possible to adopt, a priori, as a table of classification of the codes of culture, its relation to the sign to the signs and the systems of signs - and that the succession of dominant codes of culture will be, at the same time, a deeper and deeper penetration of the collective cultural conscience, in the principles that rule the systems of signs". (Ferreira, 1994/95, p. 116).

That thought transmits some ideas to the readers, as the one that says culture is a complex mechanism. Lotmam (1996) also defines the essence of culture as information and he works the semiotic of culture and the text concept. He places the problem of rapport between the culture and the fundamental categories of its transmission and conservation and the notions of language and text. His concept of cultural text talks about culture as a system of signs that somehow organizes the received information. Culture for him is formed from several languages that are part of Semiotic systems. To translate a reality in language, to transform it in a text, coded in a certain way and to introduce it in the collective memory is important for the humanity, making us think that culture is information, coding, transmission and memory.

4. Knowledge Representation in Amazonian Narratives

The Amazonian narratives are gathered by researchers of the project IFNOPAP - The imaginary in the Paraense Amazonia Popular Oral Narrative Forms, transcribed and documented in a database. The analysis of the cultural elements of the narratives is carried out by the project RESNAPAP - The Symbolic Representation of the Paraense Amazonia Popular Narratives as Language of Information. Thus, for the representation of the knowledge it was chosen the study of text of the Amazonian narratives collected from the oral tradition, in the State of Pará. The areas chosen for this study were: Abaetetuba, Belém, Bragança and Santarém. The analysis started from the narratives by means of the cultural term, considered as lexical unit of the natural language, that represents the elements of the narrator's culture, that goes on forming the cultural terminological vocabulary, good to guide us to the definitions of the terms in relation to the context of the histories. It was initially published the terminological vocabulary of the Abaetetuba, Belém and Santarém areas (Oliveira, 2001 b).

In the histories, there are the myths, the domestic and work utensils, the beliefs, the means of transportation, foods, fauna, flora and many other items that characterize the Amazonian culture. The terms removed to form those categories
are mapped, forming the terminology to specific areas, in relation to culture and space by means of geographical names, of the histories collected, taking to the integration of the knowledge for the communicative system as collective memory.

The space defined as the area of the Paraense Amazonia is formed by rivers, forests, fauna and flora as elements which are present in the imagination of the native people of the area. They even personify mythical beings. Thus, the oral histories that characterize the popular culture of the Amazonian man are replete of mythical characters that use the geographical space to make their actions. And, the narrators, this way, establish an intense identity with the place and the place with the characters. In the histories, by means of the speakers’ imagination, there are cultural terms, together with the identification of the space, forming the representation of the information through some knowledge. Some examples can be observed in the following terms from Amazonian narratives:

CURUPIRA - it is a mythical character and his habitat is the forest. The term represents the category MYTH in the space of his actions. This term is defined as a fantastic entity that inhabits and protects the forests, whose characteristic is to have the feet in a backwards position, he can assume the form of any human being to fool his victims. Depending on the histories he can appear in several situations.

For example:

“(...) O curupira deixou a mulher sair e veio no corpo da mulher(...). Criança desaparecida. Historia de Maria Rosa Leal. (Abaetetuba).

“(...) The Curupira let the woman go and came in her body (...)” The Missing Child.

“-Não é a curupira que chamam, né? (...) . Diz que um pretinho. História de Dona Flora. (Belém).

“-It isn’t the Curupira who they call, is it?” A little black boy.

“-Então, o caçador foi caçar e se perdeu nas matas. Quando foi ali, uma meia-noite, ele viu aquele barulho e disse:/- Ah! Possível. Isso pode ser curupira. (...) A curupira. História de José Travassos da Cruz. (Santarém)

“-So, the hunter went out for hunting and got lost in the woods. Around midnight, he saw a noise and said: - Ah! Possibly it’s the Curupira (...)The Curupira.

BOTO - it is a mythical character that inhabits in river and in land. As Amazonian fauna, it inhabits the river. As seductive man inhabits in land. It represents the category MYTH and the category FAUNA. Depending on the histories it can be man or woman, seductive or audacious.

For example:

“Então, o pai entendeu que a filha estava sob o encanto do boto”. O boto. Historia de Ruth Helena. (Abaetetuba)

“So, the father understood the daughter was under the boto’s spelling”. The Boto.


“It was a beautiful boy, wasn’t he? White, blond, blue-eyed. His hat was a ray. Those rays! The cane was a needle-fish. That was the Boto. Beautiful Boto.

E o boto, ele sempre aparecia , é ... pra homens também. (...) Um boto diferente. Historia de Maria Oliveira da Cunha. (Santarém)
“And the Boto, he used to appear, eh... to men too. A different Boto.

DEER - it is a term of the category fauna. It is a ruminant mammal, of ramified horns, belonging to the family Cervidae. The space that it inhabits is the varzea areas and the firm lands of the Amazonia forest. It represents the category FAUNA. It is defined as game animal and it is good for the feeding of the population that inhabits the forest. It is usually hunt at night with torches due to the darkness of the forest.

For example:

“E eu a esperar esse veado. Eu com ele, esperando o veado, e o veado nada de sair, sabe?” Caçando veado. História de Manoel Paulino da Fonseca. (Abaetetuba)

“I was waiting for that deer. He and I, waiting for the deer, and the deer didn’t want to come off, you know? Hunting deer.

MANIVA - it is a term of the category flora. The scientific term is Manihot utilissima. A plant whose roots are used in several types of flour, the tapioca and the tucupi. The leaves of the cassava, when grinded, are used to prepare the maniçoba, an exotic food very consumed in the Cirio de Nossa Senhora de Nazaré (a regional religious party) epoch. The space is the firm land for plantation. It is defined as the leaves of the cassava and it is good to elaborate a typical regional meal denominated maniçoba, that is an indigenous name.

For example:


“(...) Then he took the mandioca – the maniva – and cut it”. Stubborn.

In relation to the places mentioned in the history, many times they are identified as concrete and some others as abstract. Those places have a denomination, that is a geographical name that is understood as the proper name of places and geographical accidents. With the reading of the narratives various geographical names are identified and, for methodological reasons, they were divided into identifiers of:

a) Places. They refer to the geographical points crested or frequently used by the man for the satisfaction of his work needs, leisure, habitation, transport, such as: avenues, districts, communities, highways, municipal districts, squares, highways, streets and others. As an example, the avenue Alcindo Cacela. Avenue is the name of the place and Alcindo Cacela is what identifies the place. Name given in allusion to a historical politician and to remind the fact happened in the history.

b) Geographical Accidents. They correspond to the natural geographical points of the landscape, as: rivers, bays, islands, streams, lakes, beaches and others. For example, the Marajó Island. The island is the geographical accident and Marajó is what identifies the island.

In that way, the specialized knowledge interacts by means of the signs and of the collective cultural conscience, mapped by the Amazonian cultural terminology, for specific geographical areas of the Parelense Amazonia for the universality.

5. Final Considerations

Knowledge representation is studied, starting from popular oral narratives, informed by narrators that live in Amazonian areas, is an indicator of culture type.
It demonstrates that the semiotic is able to reveal the complexity of the social life, relationships that are established among human beings and the nature that surrounds them. The units of knowledge for the socio, geographical, cultural and linguistic study will approach culture, history, geography and communication as a semiotic product. The result aims at the integration built for the classification of the terms obtained from culture and space by means of the signs, in a great oral and popular text. The system of signs will work to help the user to understand the organization and the representation in order to retrieval the information. Therefore, in the Knowledge Era the user will be prepared to integrate the universality. The mapping of the terms in their classification will give conditions to the socio-communicative function of the text in relation to the process of communication. The collective memory comes from several narrators and from the cultural tradition to form the great text and also to be placed in the great communication net for the access of no-borders Societies.

References
Oliveira, Maria Odaisa Espinheiro de (2001b). Vocabulário terminológico cultural da Amazônia paraense: com termos culturais das áreas de Abaetetuba, Belém e Santarém. Belém, 159p. v.1
Abstract: Retrieving information from the Internet has been a difficult task for several reasons, such as the great amount of information currently accessible online. This reality has led different research areas to study filtering techniques to improve the accuracy of responses to specific requests. This paper introduces a user-friendly theoretical filtering methodology, called metafilter, based on the concept that people represent themselves and the world around them by using metaphors, which could be used for information retrieval. The ontological concept of facets as proposed by Ranganathan – Personality, Matter, Energy, Space and Time – guides our classification of metaphoric representations. A case study based on a discourse community in the Transportation field was used to illustrate our proposal.

1. Introduction

The increasing number of inscriptions currently circulating in different social contexts, largely stimulated by the advance of new technologies, has caused the adversity of a high degree of disinformation, as information-retrieval capacities lag behind the accelerating production of knowledge. Solla Price (1965) anticipated such growth in knowledge construction when he stated that academic production grows exponentially, duplicating its production every 10 years. Thinking in terms of the Internet, such possibilities are even greater. Structured as a network, this virtual communication medium allows an almost unrestricted supply of and access to data. This facility, as was pointed out by Pierre Lévy (1999, p. 13), has caused an explosive and chaotic growth of telecommunications, multiplying and accelerating the amount of available inscriptions.

Concerned with the conceptual problems that guide knowledge organization and processing, we have bent over not only the production, but especially the retrieval of available information, in consonance with studies in the area of social identity and memory. This is the framework of our work: to think about information retrieval from the Internet, including aspects of the identity representation of different social groups which use this huge database, with the purpose of allowing the filtering of relevant information.

On the one hand, metaphors are a linguistic resource that moves a given term into a sphere of meaning which was not originally related to it, with the purpose of representing the world by means of analogies. On the other hand, the set of meanings established within a group provides us a hint of the identity of such group, thus helping recognize both its more direct areas of interest and its knowledge-organization structure – and, consequently, its way of retrieving
information. If we can establish this identity, we may obtain the most relevant areas to the group and, thus, determine a possible filtering of the available information.

We assume that, in academic environments, the constituted social groups employ a given set of metaphorical representations for their object of study/work. Such representations establish an identity cohesion which allows group members to recognize themselves as such – an identity of the group itself – and to select terms and specificities relevant to the retrieved documents. The viability of this analysis served as a base for the development of an empirical framework, by means of a case study, to obtain the set of metaphorical representations a given social group – in our case, researchers in the Transportation field – employs to represent its object.

The theoretical proposition of a metaphorical filter to support information retrieval based on the thematic identity of social groups is based on the interactive view of meaning construction, and has the Nonchalance Principle (Berrendonner, 1989) as one of its theoretical backgrounds. Following, we will present a theoretical base according to which the representations guide the memory and identity of the social group, which in a way define its priorities and relevancies for data retrieval. Subsequently, we will establish an interface between metaphors and the classification of facets as proposed by Ranganathan, aiming at showing that such a relation can both evidence a group’s identity, as is illustrated by the case study, and guide the organization of the knowledge constructed and “consumed” by that social group.

2. Interactive Construction of Meaning

The conception of an interactive construction of meaning assumes a constant retro-supply among the interacting parties, so that any utterance is linked to the meaning of the previous one, and so forth. We claim that this construction is based not only on what is explicit, but also on what can be inferred from the utterances, as well as on what constitutes features of identity and memory.

2.1 Nonchalance Principle

Our proposition is based on representations as shared meanings constructed through an interactive-communicational process whose organization is ruled by Berrendonner’s (1989) Nonchalance Principle. It establishes an interface between communication and cognition according to which utterances contain little informative material, which is made possible because the speaker considers that the interlocutor already has an information background that will allow him/her to infer further specifications from a sub-specified utterance.

Such an inferential process, according to Berrendonner, is based on a set of correlated concepts that surround a nuclear and guiding concept. He names such set “conceptual constellation” and the nuclear concept “attractor”. This astronomical metaphor suggests that, in each communicational situation, sets of meanings are established in which occasionally one of the terms plays the part of gravity center of the system, allowing correlated concepts to be “activated”, ready to enter the utterance process and, consequently, the representational set of those involved in the communication.

The assumptions that support our conception of meaning take into account the conditions for the production of utterances, as they admit the occurrence of semantic shifts according to the several combinatory relationships adopted by the conceptual elements that take part in different communication situations. Based on
these principles, we assume the meaning to be constructed in an interactive situation between the one producing the utterance and the one “consuming” it.

We must also examine information transference, particularly in academic circles. It is mostly written/read and, in this sense, the practice of reading is seen as “knowing to be involved in an interaction with someone in a specific social-historical moment, and that the writer, as any other interlocutor, is using language from an established social position. To read is to be involved in a social practice” (Moita Lopes, p. 1996). This implies that the interlocutor carries a knowledge background which is necessarily assumed in the interaction process. Such background consubstantiates as a cumulative-dependent function of knowledge and experiences, being essential for meaning construction, which is not only a result of a symbolic transmission along a timeline (collective diachronic memories), but also – and above all – relates the parties involved in the communication, implying that this construction takes place in consonance with the cultural aspect of the social groups such parties belong to. To study these cultural aspects is also, in a certain way, to seek to understand the aspects of the group’s social identity and memory.

2.2 Identity and Memory

Le Goff (1996) points out that a revolution took place in the developments of memory along the 20th century, particularly after the 50’s. He attributes the technologies that appeared during that period – large calculators, for instance – a storing value that has changed the relationship between man and data storage systems. However, for the purposes of this article, what interests us most is the approach to memory as presented by Huyssen (2000, p.37), who states that “memory is always transitory, notoriously untrustworthy and prone to forgetfulness; in synthesis, it is human and social”, for we aim at considering how man establishes representations.

We consider forgetfulness as a constitutive element of memory, since memory is ruled by choices which end up determining what will remain as memory features and what will become traces that will later be forgotten. In this “game”, we assume identity and memory to be manifest by means of the discourse, since the representations are uttered by it. Pêcheux (1999, p.50) defines memory as something understood in the intercrossed meanings of a mythical memory, of a social memory embedded in practices, and of a memory constructed by historians, which leads us to consider the discursive manifestations through which such meanings are exposed.

We can therefore assume, despite the impact of technology as a factor of interference in the construction of memory, that what interests us in memory studies is its human character, prone to failure. Being an essentially human phenomenon, memory blends with identity, each one strengthening the other. Phenomena that will be either forgotten or remain memorized derive from identity features that condemn them to one situation or the other, and are materialized in the discourse by the set of metaphors selected by a social group as the one that better represents its informational interests.

To speak of interests is, somehow, to bring forth the concept of relevance. To Saracevic (1970), in the realm of information retrieval, relevance is defined by the correspondence between a document and a question, by the degree of adjustment needed between the document and the user’s previous knowledge, and by the level of signification that aims at a purpose. We can thus infer that relevance criteria are random and personalized, since they assume both the user background –
which one may attempt to level, but will never match – and the notion of purpose – a variable that strongly depends on the timeline.

3. Representations

Our goal is to present the metaphors as a viable resource for organizing knowledge aiming at information retrieval, thus evidencing their guiding role in the construction of the identity of a social group. We will also attempt to show that a set of metaphorical representations can help organizing knowledge, establishing an interface with Ranganathan’s facets.

3.1 Metaphors

As we already said, a metaphor is a figure of speech which transfers a term to a sphere of signification which is not originally related to it, seeking to represent the world by means of analogies. Lakoff and Johnson (1980) suggest that humans organize knowledge by means of structures called Idealized Cognitive Models and that categorical structures derive from this organization. This proposal assumes that mental organization occurs by means of the cultural construction of world-knowledge schemes. In order to be represented, such socio-culturally established schemes are shared by members of the social group. The novelty of this proposal was to consider categorical representation as socially constructed. Based on this model, we suggest that features of social identity and memory of a given group are expressed by its chosen socially determined sets of metaphors.

Lakoff and Johnson also introduced the concept of ontological metaphor, a cognitive model that guides the representation of man in the world. We will displace this concept from world representation and apply it to the representation of a field of knowledge, thus proposing that the recurrent occurrence of this phenomenon organizes discursive manifestations. Based on the identification of metaphorical representations in an area of expertise, we will introduce a theory of knowledge organization which helps the members of the group to retrieve information more precisely.

Assuming that the metaphors represent socially constructed categories, we will now, before presenting the case study, establish their link to knowledge representations.

3.2 Facets

Thinking about representations of meaning leads us to the father of Library Science, Shiyali Ramamrita Ranganathan (1967). We assume there is no need to reinforce his revolutionary role in the fields of classification and indexing theory. His classification view, though new and with important repercussion in classification theory as a whole, reflected the modern conception of knowledge construction, which, between the 19th and 20th century, assumed that a given discipline had precise boundaries and a conceptually delimited theoretical base.

The organizational scheme proposed by Ranganathan defines concepts that will be very useful to our analysis, especially those of entity – focused on the concrete or conceptual – and of universe – a set considered within a given context. The concept of entity is important to us because it refers to the area to be represented, so we consider that it establishes a group’s conceptual identity boundaries. To conceive any filtering mechanism is, above all, to define the nuclear concept of the area of interest. The concept of universe relates to the already mentioned referential constellation, evidencing the set of metaphorical
representations that illustrates this field of knowledge. We assume the universe of concepts to be “contained” in a group related to a nuclear concept.

Following, we will identify, by means of a case study, the set of metaphors representing the Transportation area and relate them to the five fundamental categories presented by Ranganathan – Personality, Matter, Energy, Space and Time.

4. Case Study

The case study that illustrates our analysis was performed with a research team of 28 professionals with different levels of expertise and from different regions of Brazil, devoted to Transportation studies. Our purpose was to cover a common and recurrent set of representations which worked adequately to an expressive number of professionals.

Data collection was made by means of a questionnaire sent and received by e-mail, seeking to catalogue representations of this area by analyzing the following question: “What does Transportation mean to you?” Our methodology was to examine the linguistic periods in which the professional explicitly used – or obviously intended to use – verbs to be or to represent, as in “Transportation is...” or “Transportation represents...” Such analysis revealed a recurrent use of terms comprehended by three semantic spheres: a) network / set / system; b) movement / displacement; c) means / form / proceeding.

The examinations of such data suggested that representations relative to semantic group (a) related to the Personality category, as it synthesizes the essence of this entity. Thus, the ontological metaphor of the Transportation area for our team is: “Transportation is a systemic network”.

In semantic group (b), we realized that the linguistic periods were mostly constituted by sentences employing the verb to be, but including the meaning – whether explicit or not – of verbs to make possible and to allow. Therefore, this set was related to category Matter, as it referred to the work people perform to achieve a final product, that is, its function: “Transportation is movement”. We can illustrate by quoting a reply from a researcher: “As an economical and social activity, I would say Transportation is the set or resources used to allow displacement...”

Group (c) contained metaphorical representations that related the Transportation area to category Energy. Ranganathan’s definition of this facet is that energy is manifest in spiritual, mental and physical activities which are translated into something inanimate, animate, conceptual, intellectual or intuitive, but taking humans as a reference. Since we are actually dealing with the representations humans make of their field of knowledge, we decided to use such category, because what interests us is how the linguistic community conceptualizes this phenomenon. In our case, the metaphor used was “proceeding”, as it deals with how it is done. An illustrative example is: “Transportation are the necessary proceedings to displace people, goods or information”.

Facets Space and Time, concerning Transportation, are inherent to its very essence, since they are closely related to the concept of “displacement”, and the members of this social group will hardly conceptualize ideas so close to their usual activities. According to Lakoff and Johnson (1980, 3), “In most of the little things we do every day, we simply think and act more or less automatically along certain lines” and, also, for a group’s identity and memory.
This idiosyncrasy of the Transportation area reinforces the idea that the set of metaphors establishes – or is established by – the identity of the group that employs it, determining relevancies which then will define what must – or must not – be retrieved.

5. Proposed Metafilter

To see a group’s identity as a guide for organizing produced and retrieved knowledge aims at proposing a conceptual framework that supports the creation of an information-retrieval filter to be used in a culturally globalized world. Therefore, to conceive such filter implies dealing with the varied accumulated knowledge from members of several social groups.

In order to establish a closer relationship between the theoretical discussion on communicative processes and a reflection on filtering, we insist that, in the communicative process, the only common contents within a given social or cultural group is that whose meaning (information) is considered as common contents by the members of such community, thus characterizing its identity features.

Examining the interactive process of meaning construction, we somehow relate the information retrieval process with the reception axis, and not only with the production axis, since information is associated to meaning, thus assuming a semantic understanding of the utterance. We also admit, as stated by Eco (1986), that the text generation process embeds an anticipation of the interpretation, since “decoding” a verbal message means to have not only linguistic ability but also a competence of a varying circumstantial essence, which allows for assumptions and the repression of idiosyncrasies. Such abilities imply a certain degree of subjectivity which – though it may seem a paradox – the proposed metaphor-based filter attempts to minimize.

To design the metafilter, we have attributed the metaphorical representation a similar characteristic to the one we applied to Ranganathan’s facets, as both are contextually situated within a given domain or subject. The metaphorical representations also reflect the specificities of a social group inserted in a domain or that employs it, allowing its members to recognize themselves as participants.

The metafilter conception requires two metaphorical sets for the several knowledge areas: one representing essential conditions and the other representing accessory conditions. In our studied area, the essential conditions include categories Essence, Function and Way. They are correlated to facets Personality, Matter and Energy, respectively represented by metaphors “systemic network”, “movement” and “proceeding”. The accessory conditions, for reasons already exposed, are represented by facets Space and Time, occupying generic categories called Peripheral, with no metaphorical representation.

6. Conclusion

Any classification criterion, with its subsequent relevant criteria, ultimately seeks to fulfill user expectation. In this case, our conceptual filtering proposition aims at employing metaphors that represent the categories satisfying the conditions necessary to represent a given area of knowledge, and therefore can characterize the social group’s identity. On its turn, this identity establishes the relevancies according to which the group’s informational demands are organized. Therefore,
the metaphors may support an identity representation that both allows the group to recognize itself as such and specifies the concepts used in its field of expertise.

Notes
1 Article derived from the author’s Doctorate thesis called Binômio Lingüística-Ciência da Informação: Abordagem Teórica para Elaboração de Metafiltro de Recuperação da Informação (linguistic-information science binomial: theoretical approach for the design of an information-retrieval metafilter), defended at IBICT/UFRJ (Federal University of Rio de Janeiro, Brazil), oriented by Prof. Dr. Maria Nélida Gonzáles de Gómez.

References
Using Neural Networks for Multiword Recognition in IR

Abstract: In this paper, a supervised neural network has been used to classify pairs of terms as being multiwords or non-multiwords. Classification is based on the values yielded by different estimators, currently available in literature, used as inputs for the neural network. Lists of multiwords and non-multiwords have been built to train the net. Afterward, many other pairs of terms have been classified using the trained net. Results obtained in this classification have been used to perform information retrieval tasks. Experiments show that detecting multiwords results in better performance of the IR methods.

1. Introduction

In this paper we present a new approach to solve the task of effectively detecting multiwords. A multiword is a succession of words whose sense taken as a whole differs from the sum of the senses of its single words. Thus, a multiword can be considered in fact as a new concept.

There exists a second kind of multiword composed of a set of words that complement their senses. Nevertheless, considering this kind of successions as multiwords does not provide information useful for information retrieval tasks. For this reason, this kind of multiwords have not been considered in the present work.

Multiword detection can be successfully used in many different tasks. Information Retrieval (IR) methods, for instance, use the word as the basic unit of information; thus, detecting multiwords in corpus and queries make IR systems get better results. Cross Language Information Retrieval can be also improved by detecting multiwords, given that translating word by word results in lost of information. Natural Language Processing can also be helped when multiwords are correctly detected, because it makes the text easier understanding.

The new approach presented in this paper uses neural nets to discriminate pairs of terms that really are multiwords from those that are not. We propose a well-known supervised neural network: Kohonen’s Learning Vector Quantization (LVQ), widely used for classification tasks (Kohonen, 1995, 1992). Inputs for the nets are the values yielded by estimators used in literature to perform this same task, and the output the nets provide is a class determining if that values corresponds to a multiword or a non-multiword. Nets learning is performed by training them with the values yielded by the cited estimators when they are applied to pairs of terms known to be either multiwords or non-multiwords. In order to test this new method, the network has been used to classify new pairs of terms; then, the information obtained has been used to perform some IR tasks. Experiments show that results obtained in these task are better than those obtained using only the estimators.

The rest of the paper is organized as follows: Section 2 gives a little introduction to the state of the art, briefly showing some of the currently available methods used to detect multiwords. These methods include the different estimators
that will be lately used in our method. Section 3 describes a new estimator developed for this work as well as the neural network that has been used. Section 4 shows the experiments carried out and the results obtained. Finally, Section 5 outlines some conclusions, and also future research lines.

2. A new approach

Multiword recognition has been explored by many researchers as a way to improve traditional Text Retrieval, in general with a moderate degree of success. However, David Hull and Gregory Grefenstette (Hull, 1996) show that multiword detection and correct translation largely improve the precision in a CLIR system.

Usually methods for automatizing terminological procedures have traditionally been statistical (Hull, 1996; Ballesteros, 1998), and based on the co-occurrence of each particular pair of words in the text of work or corpus. Other works (Adriani, 1999) obtain the degree of similarity between terms using the co-occurrence factor, and the standard tf*idf term weighting formula. Recently, hybrid approaches incorporating linguistic information have been developed: Diana Maynard and Sophia Ananiadou (Maynard, 2000) make use of different types of contextual information: syntactic, semantic, terminological and statistical. Nevertheless, managing different types of information must be done by integrating them in any given way. The most straightforward way is by using a linear function, although this does not mean it is the best way this problem can be faced.

For any of the features (syntactical, semantical, terminological and statistical) to be integrated to perform multiword detection, there are some well-known estimators. This paper introduces a neural network based approach that integrates terminological and statistical estimators. Multiword detection is then thought as a categorization problem where only two categories have to be managed: multiword and non-multiword. Consequently, classifying a pair of terms turns into a two step process: firstly, obtain the values yielded by the different estimators; secondly, use those values as inputs for the neural network, and obtain the class to which the pair of terms belongs. More precisely, the estimators that have been used in this work are the following:

1. Pearson's $\chi^2$. A variant of the $\chi^2$ statistic (Hull, 1996)
2. Measure the importance of co-occurrence of the elements in a set by the em metric (Ballesteros, 1998)
3. Dice similarity coefficient obtain the degree of similarity or association-relation between terms using a term association measure and the tf.idf weighting formula (Adriani, 1999).
4. The mutual information ratio, or association ratio, $\mu$ (Johansson, 1996).
5. Finally, a new estimator, a variant of Dice similarity coefficient based on the Simpson index, has been developed.

2.1. A new estimator: Simpson Similarity coefficient

Roughly, Dice index is based on the association between two terms by calculating the coefficient of the intersection of two sets and their union. Usually, this approach is convenient to estimate the correlation between words, but not always. "Bill Clinton", for instance, is a multiword, but "Bill" is a very common word, so the term frequency is very high, and "Bill" set is huge. In the other hand,
"Clinton" is not too frequent, so "Clinton" set is small. Thus, the coefficient of the intersection and the union of both sets will be small, because "Clinton" set is small. In another way, the Simpson index estimates the association between two sets by calculating the coefficient of the intersection of two sets and the smaller of them, so "Bill Clinton" will reach a high value for the Simpson coefficient, and a low value for the Dice coefficient.

\[
\text{DICE: } xy = 2 \frac{\sum_{i=1}^{n} (w'_{xi} \cdot w'_{yi})}{\sum_{i=1}^{n} w^2_{xi} + \sum_{i=1}^{n} w^2_{yi}} \\
\text{SIMPSON: } xy = 2 \frac{\sum_{i=1}^{n} (w'_{xi} \cdot w'_{yi})}{\min\left(\sum_{i=1}^{n} w^2_{xi}, \sum_{i=1}^{n} w^2_{yi}\right)}
\]

where:
- \(w_{xi}\) = the weight of term \(x\) in the document \(i\).
- \(w_{yi}\) = the weight of term \(y\) in document \(i\).
- \(w'_{xi} = w_{xi}\) if term \(y\) also occurs in document \(i\), or 0 otherwise.
- \(w'_{yi} = w_{yi}\) if term \(x\) also occurs in document \(i\), or 0 otherwise.
- \(n\) = the number of documents in the collection.

### 2.2. Neural Network approach: The LVQ algorithm

The LVQ algorithm is a classification method based on neural competitive learning, which allows to define a group of categories on the space of input data by a reinforced learning, either positive (prize) or negative (punishment). LVQ uses supervised learning to define class regions in the input data space. To this end a subset of similarly labeled codebook vectors is placed into each class region.

Given a sequence of input data, an initial group of reference vectors \(w_k\) (codebook) is selected. In each iteration, an input vector \(x_i\) is selected and the vectors \(\hat{W}\) are updated, so that they fit \(x_i\) in a better way. The LVQ algorithm works as follows:

For each class, \(k\), a weight vector \(w_k\) is associated. In each repetition, the algorithm selects an input vector, \(x_i\), and compares it with every weight vector, \(w_k\), using the euclidean distance \(\|x_i - w_k\|\), so that the winner will be the weight vector \(w_c\) nearest to \(x_i\), being \(c\) its index:

\[
\|x_i - w_c\| = \min_k \|x_i - w_k\|
\]

The classes compete between them in order to find the most similar to the input vector, so that the winner is the one with less euclidean distance regarding the input vector. Only the winner class will modify its weights using a reinforced learning algorithm, either positive or negative, depending on the classification being correct or not. Thus, if the winner class belongs to the same class of the input vector (the classification has been correct), it will increase the weights, coming slightly close to the input vector (prize). To the contrary, if the winner class is different from the input vector class (the classification has not been correct), it will decrease the weights, coming slightly far from the input vector (punishment).

Let \(x_i(t)\) be an input vector at time \(t\), and \(w_k(t)\) represents the weight vector for the class \(k\) at time \(t\). The following equation defines the basic learning process for the LVQ algorithm,
\[ w_c(t + 1) = w_c(t) + s \cdot \alpha(t) \cdot [x_i(t) - w_c(t)] \]

where \( s = 0 \), if \( k \neq c \); \( s = 1 \), if \( x_i(t) \) and \( w_c(t) \) belong to the same class; and \( s = -1 \), if they do not, and where \( \alpha(t) \) is the learning rate, being \( 0 < \alpha(t) < 1 \), a monotonically decreasing function of time. It is recommended that \( \alpha(t) \) be rather small initially, say, smaller than 0.5, and that it decrease to a given threshold, \( u \), very close to 0 (Kohonen, 1995).

The experiments showed in section 4 were carried out using the implementation described in LVQ_PAK documentation (Kohonen, 1991) with default parameters. Thus, every experiment started with the same number of codebooks per class (10 for class 0 and 10 for class 1) and the learning rate being initialized to 0.3.

3. Experiments and results

In order to train and test the neural nets, a set of samples composed of input-output pairs had to be built, every sample corresponding to a pair of terms. In one hand, input values were obtained by applying the different estimators described in section 3. In the other, every output value consisted on a single number classifying the sample as multiword or non-multiword. In ours experiment only multiwords with two relevant terms have been used, and stop words have been removed from the multiword.

Obtaining a list of multiwords was done by resorting to WordNet (Miller, 1995), a lexical database where multiwords can be found. Nevertheless, not all the pairs of terms said to be multiwords really were. For this reason, each multiword returned by WordNet was newly searched in the electronic dictionary Encarta to remove pairs of terms that, even appearing together very frequently, were not real multiwords.

Non-multiwords list (needed to train the nets) was taken from the corpus used in CLEF 2000. Pairs of terms were taken from this corpus and then searched in the list of multiwords previously described, checking that they did not appear in it. If they did not appear, they were once more searched in the electronic dictionary to assure they did not formed a multiword.

Once both multiwords and non-multiwords lists have been created, the above cited estimators were applied to them, obtaining the file with the samples to be used with the supervised network. This file was split to use 75% of the samples to train the neural network and the remaining 25% to validate it.

Los Angeles Time 1994 collection, borrowed from the English CLEF 2000 collection, was used to test the method. This collection is composed of 113,005 articles of the 1994 edition of Los Angeles Times, and 40 queries (Title + Description) with relevance judgments. The collection was indexed twice using Zprise software, with Okapi (Robertson, 2000) weighting formula. First index was created without carrying out multiword detection, while second index uses multiword detection, as depicted above.

Table 1 shows average precision reached by both methods. It shows that multiword usage improves.

<table>
<thead>
<tr>
<th>Original query set</th>
<th>Query set with multiwords detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.375</td>
<td>0.410</td>
</tr>
</tbody>
</table>

Table 1 – Average Precision
precision scarcely. Anyway, a more detailed analysis of the results leads to conclude that multiword detection is useful for IR task. Table 2 shows the precision reached by some queries, and the detected multiwords for each one.

<table>
<thead>
<tr>
<th>Query</th>
<th>Original AvgP.</th>
<th>AvgP. With Multiwords</th>
<th>Detected multiwords</th>
</tr>
</thead>
<tbody>
<tr>
<td>#7</td>
<td>0.3969</td>
<td>0.4452</td>
<td>“world soccer”</td>
</tr>
<tr>
<td>#9</td>
<td>0.1022</td>
<td>0.2027</td>
<td>“war ii” “ii war” “war rwanda” “world war”</td>
</tr>
<tr>
<td>#3</td>
<td>0.3912</td>
<td>0.3220</td>
<td>“decisions made”, “hard soft”</td>
</tr>
<tr>
<td>#32</td>
<td>0.4126</td>
<td>0.2511</td>
<td>“women priest”, “change direction”</td>
</tr>
</tbody>
</table>

Table 2 – Four detailed queries.

As Table 2 shows, query #7 gained 5% of absolute precision because “world soccer” was effectively detected as a multiword. Results were even better in query #9, in which “world war”, “war rwanda” “war ii” and “ii war” multiwords were correctly detected. As can be seen, precision obtained in this query by the new method is twice the precision obtained without multiword detection.

On the other hand, query #3 lost 7% of precision with multiword inclusion. Bigrams “decisions made” and “hard soft” are in fact non-multiwords, but the neural network method marked both of them as being such. Finally, query #32 lost 16% of precision: “women priest” and “change direction” again, are not multiwords.

4. Conclusions and future work.

This paper presents a new method to detect multiwords. This method uses the values obtained by estimators, present in literature and developed to perform this same task, as inputs for a neural net that automatically determines whether those values belong to a real multiword or simply to a pair of terms that appear together in a document.

Results show that automatic multiword detection is useful for IR. Nevertheless, the method used must get a higher accuracy, because poor detection of multiwords damages precision of the IR system. Conservative methods must be used to assess multiwords. Classifying multiwords as non-multiwords is better than recognizing too many multiwords. In other words, multiword detection must improve precision over recall.

Future lines of research include the use of new kind of neural networks, such as Radial Basis Function Nets (Broomhead, 1988; Rivas, 2001), as well as RCE (Zboril, 2000), and also unsupervised training networks as Self-Organization Maps (Kohonen, 1995).

New estimators based on semantic information can be used to improve the results. Others applications for this method must also be investigated, especially its influence in Cross Language Information Retrieval.
2. Notes
1. Encarta is available at http://www.encarta.com [2/2/2002]. Encarta has been used because it includes proper nouns that are considered to be multiwords.
2. Cross Language Evaluation Forum (CLEF) aims at promoting research and development in CLIR. For more information, see: http://www.clef-campaign.org
3. ZPrise is a software developed by NIST. It is available at http://www.itl.nist.gov/iain/894.02/works/papcrs/zp2/zp2.html [2/2/2002]

References
Eduardo Péis  
Faculty of Library and Information Science, University of Granada, Granada, Spain

Enrique Herrera-Viedma and Juan Carlos Herrera  
Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain

On the Evaluation of XML Documents Using Fuzzy Linguistic Techniques

Abstract: Recommender systems evaluate and filter the great amount of information available on the Web to assist people in their search processes. A fuzzy evaluation method of XML documents based on computing with words is presented. Given an XML document type (e.g. scientific article), we consider that its elements are not equally informative. This is indicated by the use of a DTD and defining linguistic importance attributes to the more meaningful elements of the DTD designed. Then, the evaluation method generates linguistic recommendations from linguistic evaluation judgements provided by different recommenders on meaningful elements of DTD.

1. Introduction

Finding relevant, high quality information on the World Wide Web (WWW) is a difficult task. The exponential increase in Web sites and Web documents is contributing to the fact that the Internet users are often not able to find the information they seek in a simple and timely manner. There are many publicly available search engines, but users are not necessarily satisfied with the different formats for inputting queries, speeds of retrieval, presentation formats of the retrieval results, and quality of retrieved information. Therefore, users are in need of tools to help them cope with the mass of content available on the WWW (Kobayashi and Takeda, 2000; Lawrence and Giles, 1998).

The development of standard formats for the representation of Web documents in substantially improves the quality of information substantially retrieved by search engines. The logic structure of the documents on the web can be expressed with metalanguages like XML (Goldfarb and Prescod, 1998). The eXtensible Markup Language (XML) is a simplified subset of the Standard Generalized Markup Language (SGML) intended to make it more usable for distributing materials on the Web. XML is not a markup language, as is HyperText Markup Language (HTML, which is another well-known subset of SGML), but a metalanguage that is capable of containing markup languages in the same way as SGML. The designers of XML simply took the best parts of SGML and produce something that is no less powerful than SGML, but vastly more regular and simpler to use.

Another promising direction to improve the effectiveness of search engines involves the way in which it is possible to "filter" the great amount of information...
available across the Internet. Information filtering is a name used to describe a variety of processes involving the delivery of information to people who need it. The first filtering systems developed were based on document contents. However, it is known that more effective filtering can be done by involving humans in the filtering process. This idea is supported by the collaborative filtering systems or recommender systems (Reisnick and Varian, 1997). Usually, recommendations are obtained according to a quantitative criterion, i.e., they require a critical number of distinct recommenders to be reached. On the other hand, in typical recommender systems it is assumed that people express their evaluation judgements by means of numerical values. Sometimes, however a person could have a vague knowledge about judgement valuations, yet be unable to express his/her judgements with an exact numerical value. Then, a more realistic approach may be to use linguistic assessments to express the evaluation judgements instead of numerical values, i.e., to suppose that the variables which participate in the evaluation process are assessed by linguistic terms.

The main aim of the paper is to present a fuzzy soft computing method based on computing with words (Herrera et al., 1996) for evaluating the informative quality of documents in XML format in order to generate recommendations. The recommendations are obtained from the evaluation judgements provided by a panel of selected recommenders using a computing method based on the LWA (Herrera and Herrera-Viedma, 1997) and LOWA (Herrera et al., 1996) operators. The recommendations are linguistic values that express qualitatively the informative quality of XML based documents with respect to an interest topic. With these recommendations the documents are arranged in linguistic informative categories and, in such a way, later they can be reused easily to assist another people in their search processes.

To do so, the paper is structured as follows. The XML is presented in Section 2. Section 3 is devoted to introduce the tools of computing with words. Then, the evaluation method of XML documents is defined in Section 4. Finally, several conclusions are pointed out in Section 5.

2. XML Based Documents

XML is a subset of SGML, but while SGML is mostly used for technical documentation and much less for other kinds of data, with XML it is exactly the opposite, being it more usable for distributing materials on the Web (Goldfarb and Prescod, 1998).

Therefore, as SGML, XML provides the rules for defining a markup language based on tags. It has been developed to keep up the proliferation of proprietary formats in use for electronic document processing and representation. It is a “descriptive” system that gives a declarative and machine-independent description of the document structure using codes that simply offer names to categorize and identify the parts of a document. This means that XML is a protocol devised to articulate structures of contents of documents instead of the appearance of documents.

XML allows for the creation of custom tokens and custom document structures. Each XML document and each element of an XML document is an object with its own properties. The main difference between SGML and XML is that many XML based documents don't need an DTD. In our case, for representing
the different evaluation variables we have worked with XML valid document (i.e. XML well formed document with correspondence, instead, to a DTD). A DTD serves as a template that helps to explain syntax and content of a document that is based on a specific DTD. Once a set of tokens is defined for a given document, we have to give tokens a syntactical structure. Such a structure is introduced in the form of a grammar in the DTD by means of a finite set of declarative statements delimited by angle brackets of the form:

```xml
<ELEMENT name content_model>
```

ELEMENT is a keyword specifying that an element or token of document structure is being declared; name denotes the name of the element. Each ELEMENT represents a tag denoted by name. Accordingly, content model is a name of a string of elements that defines a syntactic structure for the element name. It is specified using a regular expression style syntax where "|" stands for concatenation, "|" stands for logical or, "?" stands for zero or one occurrence, "*" stands for zero or more occurrences, and "+" stands for one or more occurrences of the preceding element. The content model of an element can be composed of the combination of content model of other elements, ASCII characters (PCDATA), binary data (NDATA), or EMPTY. The possible attributes of an element are given in an attribute list (ATTLIST) identified by the element name, followed by the name of each attribute, its type, and if it is required or not (otherwise, the default value is given). Hence, an XML document can be defined by a DTD and the text itself marked with tags described in the DTD. Tags are denoted by angle brackets (<tagname>). Tags are used to identify the beginning and ending of pieces of the document. Ending tags are specified by adding a slash before the tag name (<tagname>). Tag attributes are specified at the beginning of the elements, inside the angle brackets and after the tagname using the syntax "attname=value".

**Example 1.** The following DTD involved by XML represents the structure of a document that is a scientific article:

```xml
<!DOCTYPE article [
  <!ELEMENT article (title, authors, abstract?, introduction, body, conclusions, bibliography)>
  <!ELEMENT title (#PCDATA)>
  <!ELEMENT authors (author+)>  
  <!ELEMENT (author | abstract | introduction) (#PCDATA)>
  <!ELEMENT body (section+)>
  <!ELEMENT titleS (#PCDATA)>
  <!ELEMENT conclusions (#PCDATA)>
  <!ELEMENT bibliography (bibitem+)>
  <!ELEMENT bibitem (#PCDATA)>
 ].
```

According to this DTD, the document "article" is composed by a title, at least an author, at most an abstract, an introduction, a body, a conclusions and a bibliography. The body is made up of at least one section and each section is composed by its respective title ("titleS") and characters. The bibliography is made up of at least one bibitem. The title, each author, abstract, introduction, each section title, conclusions and each bibitem is made up of characters.
Example 2. An example of a document instance of DTD defined in Example 1 may be the following:

```xml
<?xml version="1.0" standalone="no" ?>
<!DOCTYPE article SYSTEM "article.dtd">
<title>An Introduction to the Extensible Markup Language</title>
<authors><author>Martin Bryan</author></authors>
<abstract>This article gives a very brief overview of the most commonly used components....</abstract>
<introduction>XML was not designed to be a standardized way of coding text: in fact....</introduction>
<body>
<section><title>What is XML?</title>XML is subset of the Standard Generalized Markup Language (SGML) defined in ISO standard 8879:1986 that....</section>
<section><title>The components of XML</title>XML is based on the concept of documents composed of a series of ...</section>
<conclusions>By storing data in the clearly defined format provided by XML you can ....</conclusions>
</body>
</article>

3. Tools of Computing with Words

A fuzzy linguistic approach is an approximate technique appropriate to deal with qualitative aspects of problems (Zadeh, 1975). An ordinal fuzzy linguistic approach (Herrera et al., 1996) is a very useful kind of fuzzy linguistic approach, and used for modelling the processes of computing with words and the linguistic aspects in the problems. It is defined by considering a finite and totally ordered label set, \( S = \{ s_i, i \in H = 0, \ldots, T \} \), \( s_i \geq s_j \) if \( i \leq j \), in the usual sense, and with odd cardinality (7 or 9 labels). The mid term represents an assessment of "approximately 0.5" and the rest of the terms are placed symmetrically around it. The semantics of the linguistic term set are established from the ordered structure of the term set by considering that each linguistic term for the pair \((s_i, s_j)\) is equally informative.

In any linguistic approach we need management operators of linguistic information. An advantage of the ordinal fuzzy linguistic approach is the simplicity and quickness of its computational model for computing with words. It is based on the symbolic computation (Herrera and Herrera-Viedma, 1997; Herrera et al., 1996). This technique acts by direct computation on labels, taking into account the order of such linguistic assessments in the ordered structure of linguistic terms. This symbolic tool seems natural when using the fuzzy linguistic approach, because the
linguistic assessments are simply approximations which are given and handled when it is impossible or unnecessary to obtain more accurate values.

Usually, the ordinal fuzzy linguistic model for computing with words is defined by establishing i) a negation operator, ii) comparison operators based on the ordered structure of linguistic terms, and iii) adequate aggregation operators of ordinal fuzzy linguistic information.

1. There is the negation operator: \( \text{Neg}(s_i) = s_j \) with \( j = T - i \).

2. Maximization operator: \( \text{MAX}(s_i, s_j) = s_i \) if \( s_i \geq s_j \).

3. Minimization operator: \( \text{MIN}(s_i, s_j) = s_i \) if \( s_j \leq s_i \).

In the following subsections, to complete the ordinal linguistic computational model we present two aggregation operators that we shall use to define the evaluation method of XML documents.

3.2. The LOWA Operator

The Linguistic Ordered Weighted Averaging (LOWA) is an aggregation operator of ordinal linguistic values based on symbolic computation (Herrera et al., 1996). It is used to aggregate non-weighted ordinal linguistic information, i.e., linguistic information values with equal importance.

Definition 1. Let \( A = \{ a_1, \ldots, a_m \} \) be a set of labels to be aggregated, then the LOWA operator, \( \Phi \), is defined as \( \Phi(a_1, \ldots, a_m) = W \cdot B^T = C^m \{ w_h, b_h, k = 1, \ldots, m \} \), where \( W = [w_1, \ldots, w_m] \), is a weighting vector, such that, \( w_i \in [0, 1] \) and \( \sum w_i = 1 \). \( b_h = w_k / (\sum_h w_k) \), \( h = 2, \ldots, m \), and \( B = \{ b_1, \ldots, b_m \} \) is a vector associated to \( A \), such that, \( B = \sigma(A) \) = \{ \( a_{\sigma(1)}, \ldots, a_{\sigma(m)} \) \}, where, \( a_{\sigma(i)} \leq a_{\sigma(j)} \) \( \forall i \leq j \), with \( \sigma \) being a permutation over the set of labels \( A \). \( C^m \) is the convex combination operator of \( m \) labels and if \( m = 2 \), then it is defined as \( C^2 \{ w_h, b_h, i = 1, 2 \} = w_i \Theta s_i \Theta (1 - w_i) \Theta s_1 = s_k \), such that \( k = \min \{ T, i + \text{round} \left( w_i \cdot (j - i) \right) \} \), \( s_i \) \( s_j \) \( \in S \), \( j \geq i \), being "round" the usual round operation, and \( b_1 = s_j, b_2 = s_i \), if \( w_j = 1 \) and \( w_i = 0 \) with \( i \neq j \) \( \forall i \), then \( C^m \{ w_h, b_h, i = 1, \ldots, m \} = b_j \).

The LOWA operator is an "or-and" operator (Herrera et al., 1996). This property allows that the LOWA operator carries out a soft computing in the modelling of MAX and MIN linguistic operators. In order to classify OWA operators in regard to their localisation between and or, Yager (Yager, 1988) introduced a measure of orness, associated with any vector \( W \) as follows

\[
\text{orness}(W) = \frac{1}{m - 1} \sum_{k=1}^{m} (m - k)w_k.
\]

An important question of the LOWA operator is the determination of \( W \). A possible solution consists of representing the concept of fuzzy majority by means of the weights of \( W \), using a non-decreasing proportional fuzzy linguistic quantifier (Zadeh, 1983), \( Q \), in its computation (Yager, 1988): \( w_i = Q(l/m) - Q((i-1)/m) \), \( i = 1, \ldots, m \), being the membership function of \( Q \).
With \( a, b, r \in [0,1] \), some examples of non-decreasing proportional fuzzy linguistic quantifier are "most" \((0.3, 0.8)\), "at least half" \((0, 0.5)\) and "as many as possible" \((0.5, 1)\).

3.3. The LWA Operator

The Linguistic Weighted Averaging (LWA) operator (Herrera and Herrera-Viedma, 1997) is another important aggregation operator which is based on the LOWA operator. It is defined to aggregate weighted ordinal linguistic information, i.e., linguistic information values with not equal importance.

**Definition 2.** The aggregation of a set of weighted linguistic opinions, \( \{(c_i, a_i), \ldots, (c_m, a_m)\} \), according to the LWA operator \( I_T \) is defined as

\[
I_T[(c_1, a_1), \ldots, (c_m, a_m)] = \Phi(h(c_1, a_1), \ldots, h(c_m, a_m)),
\]

where \( a_i \) represents the weighted opinion, \( c_i \) the importance degree of \( a_i \), and \( h \) is the transformation function defined depending on the weighting vector \( W \) assumed for the LOWA operator \( \Phi \) such that,\( h = \min(a_1, a_2) \) if orness \((W) \geq 0.5\), and \( h = \max(\neg(c_1), a_2) \) if orness \((W) < 0.5\).

4. Evaluating XML Based Documents for Generating Recommendations

Suppose that we want to generate a recommendation database for qualifying the information of a set of valid XML based documents, \( \{d_1, \ldots, d_n\} \), with the same DTD. These documents can be evaluated from a set of different areas of interest, \( \{A_1, \ldots, A_q\} \). Consider an evaluation scheme composed by a finite number of elements of DTD, \( \{p_1, \ldots, p_n\} \), which will be evaluated in each document \( d_k \) by a panel of recommenders or referees \( \{e_1, \ldots, e_m\} \). We assume that each component of that evaluation scheme presents a distinct informative role. This is modeled by assigning to each \( p_j \) a relative linguistic importance degree \( I(p_j) \) supported by the linguistic variable "Importance" defined as in Section 2, i.e., \( p_j \in S = \{s_1, \ldots, s_T\} \). Each importance degree \( I(p_j) \) is a measure of the relative importance of element \( p_j \) with respect to others existing in the evaluation scheme. We propose to include these relative linguistic importance degrees in the DTD. This can be done easily by defining in the DTD an attribute of importance "rank" for each component of evaluation scheme using the XML syntax.

Let \( e^{\mu}_{ijkt} \) be a linguistic evaluation judgement provided by the recommender \( e_k \) measuring the informative quality or significance of element \( p_j \) of document \( d_i \) with respect to the area of interest \( A_t \). Consider that \( e^{\mu}_{ijkt} \) is supported by the linguistic variable "Significance", which uses the same label set associated to "Importance", but with a different interpretation, i.e., \( e^{\mu}_{ijkt} \in S \). Then, the evaluation procedure of a XML based document \( d_i \) obtains a recommendation \( r^{\mu}_{ijkt} \), \( r^{\mu}_{ijkt} \in S \). (It is also supported by the linguistic variable "Significance") using evaluation method based on the LWA and LOWA operators which is composed by the following steps:

1. Capture the topic of interest \( A_t \), the linguistic importance degrees of evaluation scheme fixed in the DTD \( \{I(p_1), \ldots, I(p_n)\} \), and all the evaluation judgements provided by the panel of recommenders \( \{e^{\mu}_{ijkt}, j=1,\ldots,n, k=1,\ldots,m\} \).
2. Calculate for each $e_k$ his/her individual recommendation $r_{e_k}^i$ by means of the LWA operator as

$$r_{e_k}^i = \prod \left( I(p_i), e_{e_k}^i \right) = \Phi \left( h(I(p_i), e_{e_k}^i) \right).$$

Therefore, $r_{e_k}^i$ is a significance measure that represents the informative quality of $d_i$ with respect to topic $A_i$ according to the $Q$ evaluation judgements provided by $e_k$, being $Q$ the linguistic quantifier used to compute the weighting vector of $Q$.

3. Calculate the global recommendation $r_i$ by means of $\Phi$ guided by the fuzzy majority concept represented by the linguistic quantifier $Q$ as $r_i = \Phi \left( r_{e_1}^i, ..., r_{e_m}^i \right)$. Then, $r_i$ is a significance measure that represents the informative quality of $d_i$ with respect to $A_i$ according to the $Q$ evaluation judgements provided by the $Q$ recommenders. $r_i$ represents the linguistic informative category of $d_i$ with respect to $A_i$.

4. Store the recommendation $r_i$ in a recipient in order to assist users in their later search processes.

5. Conclusions

In this paper, we have presented a fuzzy linguistic evaluation method to characterize the information contained in XML based documents. The method generates linguistic recommendations for structured documents by taking into account the fuzzy majority of linguistic evaluation judgements provided by different recommenders to evaluate the informative quality of the more meaningful component of DTD. The use of fuzzy linguistic modeling facilitates the activity of the filtering systems due to that the user-system interaction is more user-friendly.

References


Artificial Neural Networks Applied to Information Retrieval

Abstract: Connectionist models or neural networks are a type of AI technique that is based on small interconnected processing nodes which yield an overall behaviour that is intelligent. They have a very broad utility. In IR, they have been used in filtering information, query expansion, relevance feedback, clustering terms or documents, the topological organization of documents, labeling groups of documents, interface design, reduction of document dimension, the classification of the terms in a brainstorming session, etc.

The present work is a fairly exhaustive study and classification of the application of this type of technique to IR. For this purpose, we focus on the main publications in the area of IR and neural networks, as well as on some applications of our own design.

1. Introduction

Throughout history, the desire to improve the results of information retrieval (IR) has led to progress from systems based on inverted indices to those based on the vector space model, the fuzzy model, etc. Techniques of classification or clustering, relevance feedback, etc., and various interfaces have been used to facilitate interaction with the user, etc. We are all aware, however, of the deficiencies of IR, such as the difficulty in expressing the query, information overload, the disarray of the World-Wide Web, etc.

In an attempt to overcome some of these problems, techniques of Artificial Intelligence (AI) are currently being used as part of, or conjointly with, other IR techniques. An example is the use of Expert Systems in query design, the creation of knowledge bases with the relationship between terms specified, genetic algorithms for query optimization (López-Pujalte et al. 2002; Cordón et al. 1999), self-organizing maps for document classification, Hopfield networks for the suggestion of terms in the form of an implicit thesaurus, etc.

Connectionist models or neural networks are a type of AI technique that is based on small interconnected processing nodes that yield an overall behaviour which is intelligent. Generically they can be said to be trainable to give a response to different inputs. They are characterized by their neurons or processing nodes, by the connections between these nodes, and by the learning algorithms that they use. By varying these characteristics, one obtains networks of widely differing
behaviour and usefulness, from the generic type that provides a response to inputs as mentioned above, to classification, clustering, topological organization, etc.

2. Applications in Information Retrieval

The applications of artificial neural networks to IR can be classified into eight categories:

2.1 Complete IR processing

Some workers see the process of document retrieval as a three-layer network: the query is the input layer, the retrieved documents are the output layer, and the hidden layer in some fashion represents the indexing terms. In the opinion of Doszkocs et al. (1990), this had been defined implicitly by certain authors [Bush 1945]. Belew (1989) was one of the first to outline a network of this type, and Kwok (1989) used it with a modified Hebbian learning algorithm.

Wilkinson & Hingston (1992) used the cosine measure in a similar model for IR. Recently, Cortez et al. (1995) implemented a system consisting of a network of this type with a backpropagation learning law and a symbolic learning system in an attempt to solve the two prime problems of Boolean systems - the lack of precision and poor specification of the queries, and the incomplete and inconsistent indexing (Larson 1992).

Brachman & McGuiness (1988) use a Boltzmann Machine to perform a conceptual retrieval by means of a process of simulated annealing. The result was satisfactory, although some aspects relative to the annealing procedure and to fitting the weights need to be reconsidered.

2.2 Query expansion and the suggestion of terms from thesauri or lists

Artificial neural networks have been used on various occasions for query expansion. One example is the Neurodoc project (Lelu 1991), in which the entire browsing process is represented as a three-layer neural network, with queries being expanded by a feedback mechanism.

Hopfield networks have been used to aid users with the current techniques, memorizing or generating thesauri, lists of headings, etc., so as subsequently to suggest terms and refine the query. Such automatically generated thesauri or concept space approaches (Chen, et al. 1997) are in great demand for databases with a very specific scientific content.

2.3 Information filtering

Jennings and Higuchi (1992) have designed a neural network that models users, and which is used to accept or reject Usenet News. This network consists of nodes (representing words) with interconnections (reproducing the relationships between terms). The nodes evolve according to the articles that are read and rejected, i.e., the nodes representing words that are present in the read articles will increase in energy, and the contrary will occur for those which represent words in the rejected articles. Likewise, the relationships between terms will increase in weight if those words appear in the same articles. It allows the user to add keywords at the time of retrieval, thus increasing their weight, and to vary the threshold, thereby changing the precision and recall that is required, in order to take into account the user's immediate interests. It also incorporates a certain decay of the weights with time - a type of "forgetting". When an article arrives, its energy is calculated (as a function of the terms it contains), and if it surpasses the threshold it
is retrieved. This user model varies gradually with time, according to the articles read and rejected.

2.4 Assistance to the user

Kantor (1993) uses an adaptive neural network in the program ANLI (Adaptive Network Library Interface) to incorporate recommendations made by the users. One can browse using a hypertext tool contained in the interface. This system allows users to work co-operatively by sharing information through computers.

Meghabghab (1994) uses neural networks in the design of INN, which substitutes for an expert in negotiating with the user in carrying out searches. It uses a network called ACN (Attentional Connectionist Network) to separate concepts, agents, and relationships, a knowledge base, and analogue reasoning.

2.5 Clustering

These networks are also beginning to be used for clustering. To this end, MacLeod & Robertson (1991) use a neural network mainly based on ART (although with some modifications). They evaluate its performance with the Keen and Cranfield collections, and obtain results that are similar to hierarchical algorithms in efficacy and somewhat better in efficiency. It has to be said, however, that their document representation was very poor (almost binary), and they did not use the possibility of parallel processing, although even so their results are considered satisfactory. Later, Muñoz (1994) used a fuzzy ART and c-mean hybrid algorithm to perform a fuzzy clustering of documents. Also, with the aim of automatically generating thesauri, there have been applications to clustering terms Oakes and Taylor 1990].

2.6 Reduction of the dimension

Backpropagation networks have been used to reduce the dimensions of document vector spaces (Muñoz 1994). This can be done with a network design consisting of very few hidden units, although it has the drawback of a computationally costly learning process.

We recently applied Kohonen’s algorithm with a fuzzification module for dimension reduction (Guerrero et al. 2001). We used a SOM for the topological organization of the database terms, and the fuzzification module to obtain degrees of membership which allow each term and each document to be characterized. This method allows one to make as large a reduction of the dimension as one wishes, as well as endowing the resulting components with a meaning that is the measure of the membership of the document to the corresponding cluster of terms. This cluster and the corresponding component may be regarded as being labeled by the terms which have the greatest degree of membership to that cluster.

2.7 Classifying electronic meeting output

Orwig, Chen and Nunamaker (1997) have used maps of this type to process and classify the remarks made in a "brainstorming" session. The evaluation was compared to a Hopfield algorithm (also a neural algorithm) (Chen et al. 1994) and to a human expert. The Kohonen network was found to be slightly inferior to the human expert (only with respect to precision). It performed its task, however, in six minutes, while the human expert took more than forty.

2.8 Topological organization

The most active field of application of neural networks to IR is in generating topological maps using Kohonen networks (Moya, Herrero & Guerrero 1998. One of the first applications of these maps in text processing was to create topological representations of the semantic relationships between words from their contexts.
(Ritter & Kohonen 1989). A similar approach was taken in the Word Category Map [Honkela et al. 1995]. A vector, divided into three equal parts of 90 components each, is assigned to the occurrence of each term. The central part corresponds to the encoding of the term in question, the first part to the encoding of the preceding term, and the third to the encoding of the following term. This really novel form of including a small context in the representation of the terms and the neural network gives rise to a map in which nouns cluster in one zone, verbs in another, etc. Because of the structure of the language, the good results obtained with English do not necessarily carry over to other languages.

WEBSOM [Kohonen et al. 1999; Lagus et al. 1999; Lagus & Kaski 1999] is a system aimed at classifying great numbers of documents, and is being applied to the Web. The documents are represented on the basis of the output from the WCM, and those vectors are used to train a Kohonen network which organizes them by topic in two dimensions. It has a well thought out graphical interface in which the intensity of colour indicates the different document densities. The representation of the documents is particularly noteworthy, since as it has few components it allows a great number of documents to be processed at little computational cost.

Xia Lin of the University of Kentucky uses these maps to generate an output display of a given document collection [Lin 1997]. The documents are represented according to the terms they contain, and they are used to train the Kohonen network. After training, each neuron is assigned the term closest to its weight vector. Hence one finally has the documents classified on a surface divided into a set of zones labeled by terms.

Chen et al. [1998a] also use a very similar system to create an information access interface, and evaluate it through a series of users who compare it with a classical system. Similar too is the Neurolsoc system developed for access to the ISOC database [Moya et al. 1999], although now the document representation is different, being based on the co-occurrence of different descriptors.

The results obtained with a very similar system have been compared with the classification given by the descriptors assigned to the documents [Guerrero-Bote et al. 2001]. The same system is used for clustering and for the topological representation of terms and documents [Guerrero-Bote 1997].

3. Conclusion and outlook

Some authors see neural networks as a general framework in which other techniques can be applied. Doszkocs et al. [1990] consider that all connectionist models can be seen as input output classification systems. Likewise, document clustering can be seen as a classification of the document space in the same space, thesauri generation as a classification of the space in terms of itself, and indexing as a classification of document space in the space of terms.

For these reasons, the most active and promising fields of application are clustering and topological organization, which may be used for retrieval itself, to refine searches, or as an interface to represent the relationships between the documents that have been retrieved.

There also seems to have been little exploration of the field of dimension reduction, which, although computationally costly, can be performed offline, allowing, in production, spaces of lower dimensions to be dealt with. This would seem to be an invaluable asset given the size and complexity of today's databases.
References


Fuzzy Logic for Measuring Information Retrieval Effectiveness

Abstract: We present a new fuzzy extension of the classical effectiveness measures of information retrieval by a new way to calculate the relative cardinality of a fuzzy set. Previous approaches using Zadeh's cardinality are compared to our new approach in an experimental stage. The experiments have been carried out with a genetic algorithm where the fitness function to optimize is a combination of the fuzzy recall and fuzzy precision measures. Results included at the end of the paper show the goodness of our proposal.

1. Introduction

Fuzzy logic has been widely utilized in different stages of the retrieval process. The representation of terms in documents and the queries with fuzzy grade in the matching task has generated the use of effectiveness measures using fuzzy logic.

The extension of the classical measures to the fuzzy case has been due to two main reasons:

- On the one hand, this extension comes from the handling of fuzzy value in the document representation by terms, since the cardinal calculus must be carried out with fuzzy sets. Therefore, the use of fuzzy cardinals is fundamental to deal with this type of representation.

- On the other hand, measures based on fuzzy logic give the flexibility that classical measures lack, since retrieved documents are counted as relevant or non relevant without any grade in the classical case.

As is well known, the most used measures in the information retrieval framework are the recall and precision. Since the first extension of these measures to the fuzzy case in 1981, there have been several approaches in the literature in order to improve the behaviour of these measures to be used in Fuzzy Information Retrieval Systems (Gedeon and Koczy, 1995), (Sanchez y Pierre, 1994), (Martin-Bautista, 2000). These approaches can be mainly differentiated by the way the extension of the classical measures to their fuzzy extension is proposed. Concretely, the two first approaches, together with the original proposal from Kraft and Buell are based in the use of the Zadeh's relative fuzzy cardinality to determine the number of documents belonging to a certain set in terms of relevance. However, in the third proposal, a different approach using a new way to calculate the fuzzy cardinality presented by Delgado, Sánchez y Vila en (Delgado et al., 2000) is used. We present in this work a new extension of the effectiveness measures for information retrieval based on fuzzy logic using this cardinality, which is based on the evaluation of quantified sentences.
2. Information Retrieval Measures with Fuzzy Logic

The performance criterion considered in the evaluation of an information retrieval system is that the query answer retrieved by the system should correspond to the user preferences, that is, the documents should be evaluated with the same grade of relevance that the user would do with all the documents in the collection.

The extension of the relevance-based effectiveness measures (Salton and McGill, 1983) consists of the transformation of the cardinalities of the sets into fuzzy cardinalities. The recall $p$ and precision $\psi$ measures in the crisp case are defined as shown in (1), where $\Omega_R$ and $\Omega_L$ represent the subsets (of $\Omega=\{\omega_1, K, \omega_n\}$) of the documents retrieved and the relevant documents, respectively.

\[
\rho = \frac{|\Omega_R \cap \Omega_L|}{\Omega_L} \quad \psi = \frac{|\Omega_R \cap \Omega_L|}{\Omega_R} \quad (1)
\]

Let $\Omega=\{\omega_1, ..., \omega_n\}$ be the document set queried by a query $Q$. The user evaluation of $\Omega$ with respect to $Q$ is characterized by a fuzzy subset of $\Omega$, $E_{\Omega,Q} = \beta_1/\omega_1 + ... + \beta_n/\omega_n$, where $\beta_i = \mu_\Omega(\omega_i)$ is the user's evaluation of the degree to which the object $\omega_i$ satisfies the query $Q$. The system answer to $Q$ over $\Omega$ is also modeled by a fuzzy subset of $\Omega$; let us denote this subset $S_{\Omega,Q} = a_1/\omega_1 + ... + a_n/\omega_n$. In order to get a perfect information retrieval system, the ranking given by the system must be the same as the ranking that the user gives to the same set of documents, that is, $S_{\Omega,Q} = E_{\Omega,Q}$ for all $i$, where $S_{\Omega,Q}^{(a_i)}$ and $E_{\Omega,Q}^{(a_i)}$ are respectively, the $a_i$-cut of $S_{\Omega,Q}$ and the $a_i$-cut of $E_{\Omega,Q}$.

The fuzzy extension of (1), applying the sigma count as the (scalar) cardinality (Zadeh, 1983), is given by:

\[
\tilde{\rho} = \frac{\sum_{\alpha \in \Omega \cap \Omega_L}(\mu_{\Omega,R}(\omega_i) \wedge \mu_{\Omega,L}(\omega_i))}{\sum_{\alpha \in \Omega}(\mu_{\Omega,L}(\omega_i))} \quad \tilde{\psi} = \frac{\sum_{\alpha \in \Omega \cap \Omega_L}(\mu_{\Omega,R}(\omega_i) \wedge \mu_{\Omega,L}(\omega_i))}{\sum_{\alpha \in \Omega}(\mu_{\Omega,L}(\omega_i))} \quad (2)
\]

Several approaches have been presented in the literature concerning the use of fuzzy logic in information retrieval. As we commented above, a fuzzy representation of the terms is needed for these fuzzy models. Such a representation depends on the retrieval system considered. In the following, we explain the main approaches to fuzzy measures.

2.1. Buell and Kraft approach

These authors made one of the first extensions of Boolean retrieval systems with the fuzzy model (Buell and Kraft, 1981). Their fuzzy representation was based in a fuzzy membership function of "aboutness" of a term to a document. The performance measurement of the model is a generalization of the recall and precision measures, calculated from two rankings of evaluations $S$ and $E$ in an analogous manner as the crisp case:

\[
\rho = \frac{|S \cap E|}{|E|} \quad \psi = \frac{|S \cap E|}{|S|} \quad (3)
\]
where \( S \) and \( E \) may be interpreted as the system's and expert's evaluation, respectively.

Using Zadeh's cardinality, these expressions are transformed into the fuzzy equivalent ones:

\[
\tilde{\rho} = \frac{\sum_{i=1}^{n} (\beta_i \land \alpha_i)}{\sum_{i=1}^{n} \beta_i} \quad \tilde{\psi} = \frac{\sum_{i=1}^{n} (\beta_i \land \alpha_i)}{\sum_{i=1}^{n} \alpha_i}
\] (4)

### 2.2. Sanchez and Pierre approach

In this approach (Sanchez and Pierre, 1994), a fuzzy extension of the term weighting is considered for Boolean retrieval systems. In this model, terms have two fuzzy values associated. On the one hand, the value of "aboutness", meaning the degree in which the term is related to a document. On the other hand, a fuzzy query weight is associated to a term and coded into chromosomes. This weight means the relative importance of the term in the query. The fitness function to evaluate is a combination of fuzzy precision and fuzzy recall. For the output, documents are retrieved according to a threshold \( T \) in answer to the query considered. The fuzzy recall and precision measures are defined basing on the extension from cardinalities to fuzzy cardinalities of the classical measures, as is presented in (5).

\[
\tilde{\rho} = \frac{\sum_{\omega_i \in \Omega} \mu_{\omega_i} (\omega_i) \land \mu_{\omega_i} (\omega_i)}{\sum_{\omega_i \in \Omega} \mu_{\omega_i} (\omega_i)} \quad \tilde{\psi} = \frac{\sum_{\omega_i \in \Omega} \mu_{\omega_i} (\omega_i) \land \mu_{\omega_i} (\omega_i)}{|\Omega_R|}
\] (5)

Although these two approaches are based on Zadeh's cardinality, there is a difference between them. In the former, all the fuzzy values of the documents in a ranking are considered. In the latter, a relevance threshold \( \theta \) is fixed, but the cardinality of the denominator in the case of the fuzzy precision is calculated as the cardinality of a crisp set of the retrieved documents above the value \( \theta \).

The expressions of recall and precision (4) measure the relative cardinality of two sets. The relative cardinality of one set \( A \) with respect to another set \( B \) defined on the same domain is the percentage of objects of \( B \) that are in \( A \). This concept has been generalized to the fuzzy case in several ways. The most used measure is Zadeh's (Zadeh 1975), which is based on the \( \Sigma \)-count measure introduced by (De Luca and Termini, 1972). The \( \Sigma \)-counts measure the energy of a fuzzy set (i.e., the total amount of membership to the set), rather than the integer number of objects in the fuzzy set (De Luca and Termini 1972). In fact, it is known that sometimes \( \Sigma \)-counts provide a high value for the cardinality of a fuzzy set due to an accumulation of small membership degrees. This is counterintuitive. When measuring relative cardinalities the situation can be worst. More details are given in (Delgado et al., 1999), where it is claimed that the only possible cardinalities of a fuzzy set are the cardinalities of its \( \alpha \)-cuts.
3. Fuzzy Measures with Delgado, Sánchez and Vila fuzzy relative cardinality

The extension of the measures presented is based on the definition of new measures of the fuzzy relative cardinality of fuzzy sets (Delgado et al. 1999) and the evaluation of type II quantified sentences (Zadeh 1983) by means of a new method that was introduced in (Delgado et al. 2000). This last approach has been successfully employed in data mining applications to obtain conditional evidences involving fuzzy items (Sánchez 1999).

3.1. Fuzzy Recall

The expression of fuzzy recall with the mentioned cardinality calculus is as follows:

\[ \rho = \text{DSV}(S/E) = \sum_{\alpha \in \Delta(E/S)} \left( (\alpha_i - \alpha_{i+1}) \frac{|S \cap E|^{(\alpha)}}{|E|^{(\alpha)}} \right) \]  

(6)

where \( \Delta(E/S) = \Lambda(S \cap E) \cup \Lambda(E) = \{\alpha_1, \alpha_2, \ldots, \alpha_q\} \), \( q \in \{1, \ldots, 2n\} \) being the cardinality of the set \( \Delta(S/E) \), and with \( 1 = \alpha_1 > \alpha_2 > \ldots > \alpha_q = 0 \). \( \Lambda(F) = \{F(x) \forall x \in X, F(x) > 0\} \) being the level set of \( F \), \( F^{(\alpha)} \) being the \( \alpha \)-cut of \( F \), and \( \cdot \) being the classical set cardinality. The set \( E \) is assumed to be normalized before the recall. If not, \( E \) is normalized (i.e. each degree is divided by a factor equal to \( \max\{E(x) \mid x \in X\} \) before the precision is calculated, and the same normalization factor is applied to \( S \cap E \).

3.2. Precision

In an analogous way, the precision is calculated as in (6).

\[ \psi = \text{DSV}(E/S) = \sum_{\alpha \in \Delta(E/S)} \left( (\alpha_i - \alpha_{i+1}) \frac{|S \cap E|^{(\alpha)}}{|S|^{(\alpha)}} \right) \]  

(7)

where \( \Delta(E/S) = \Lambda(S \cap E) \cup \Lambda(E) = \{\alpha_1, \alpha_2, \ldots, \alpha_p\} \), \( p \in \{1, \ldots, 2n\} \) being the cardinality of the set \( \Delta(E/S) \), and with \( 1 = \alpha_1 > \alpha_2 > \ldots > \alpha_p = 0 \). In this case, the set \( S \) is assumed to be normalized. If not, \( S \) is normalized before the precision is calculated, and normalization factor is applied to \( S \cap E \).

Both expressions (6) and (7) can be obtained in \( O(n) \) (n being the number of documents) if we consider only a fixed set of levels \( \Delta(E/S) \), (Sánchez, 1999).

4. Experimental Stage

4.1. Description of the System

For the experiments, we have consider the genetic model oriented to documents presented in (Martin-Bautista, 2000), where we start from a set of documents \( \Omega = \{D_1, \ldots, D_K, D_{K+1}, \ldots, D_m\} \) ordered decreasingly and evaluated by the user, being \( u_i \in [0,1] \) the user's evaluation of the document \( D_i, i = 1, \ldots, m \), that is, the degree that the user consider a document relevant. The set \( \Omega \) is divided into
two subsets $\Omega_R = \{D_1, \ldots, D_K\}$ and $\Omega_{NR} = \{D_{K+1}, \ldots, D_m\}$, containing the relevant documents and the non-relevant documents, respectively.

The relevance threshold from which we consider a document relevant is $\alpha \in [0,1]$, initially fixed to 0.5. We shall assume that an evaluation $u > 0.5$ indicates a good document, while $u \leq 0.5$ indicates a bad document. The fitness function to optimize is a combination of fuzzy recall and fuzzy precision as follows, called fuzzy recall-precision ($\tau$) given by:

$$\tau = \rho \psi \nu,$$

where $\rho$ is the fuzzy recall, and $\psi$ is the fuzzy precision defined in the former section and $\nu_1$, $\nu_2$ are the importance weights of fuzzy recall and fuzzy precision, respectively.

Let $T = \{t_1, t_n\}$ be the set of terms extracted from document base $\Omega$, and $x_{ij}$ the relative (normalized) frequency of term $t_j$ in document $D_i$. The estimation of the expected value of $x_{ij}$ in good and bad documents is given by the weighed average of the relative occurrence frequency of a symbol $t_j$ in the collection. The knowledge of the system about the user preferences is kept in the population of a Genetic Algorithm. A gene in a chromosome is defined by a term and a fuzzy number $\tilde{h}$ of occurrences of the term in documents belonging to the class of documents that satisfy the user's information need. The details of this fuzzy representation can be seen in (Martín-Bautista, 2000).

4.1.1. Genetic Components

The genetic components have been determined by preliminary tests. The selection scheme for probabilities is the inverted linear ordering (Bäck, 1992). The selection mechanism is the universal stochastic sample (Goldberg, 1989), with an elitism model. To carry out the crossover, we have considered the one-point crossover operator, with a probability of 0.6. As for the mutation, we choose by a standard mechanism a gene to mutate. Finally, the size of the population is 80 chromosomes and the chromosome length is 10. We have considered 1000 generations for each run of the algorithm, calculating the average of three runs with different random seeds to get the final results. The importance weights of the fuzzy recall ($\nu_1$) and fuzzy precision ($\nu_2$) are set as $\nu_1 = 0.67$ and $\nu_2 = 1$.

4.2. Document Sets

We have considered two different examples to carry out the experimental stage. The first example is a collection of documents corresponding to the following query in the INSPEC database of Jul-Sep 1998: "Information Retrieval and Classification". The number of documents retrieved was 22, and the number of different terms extracted (after removing stop-list words and stemming) is 616. The number of total terms is 1,074. Let suppose that the user's information needs are oriented to those documents (in the collection of the 22 documents retrieved previously), regarding topics such as Web and Internet, to which the user will give the higher evaluation. Initially, the documents are evaluated by an expert, where there are 8 relevant documents and 14 non-relevant ones, with the relevance threshold set to 0.5.

The second example is a collection of 100 documents, corresponding to the query "Information and Retrieval" to the INSPEC database of 1999-2000. The number of terms extracted for these documents is 11,210, where 1641 terms are
different. The preferences of the user in the framework of this query are located in the field of Genetic Algorithms. As in the first example, the user will give the highest relevance to the most preferred documents. In this case, the number of relevant documents is 13 and there are 87 non-relevant documents.

4.3. Classification Errors

Additionally to the effectiveness information retrieval measures, we can define classification errors to check the goodness of the measures, and compare among the fuzzy measures using Zadeh’s relative cardinality and Delgado, Sánchez and Vila’s We can compute a global error comparing the relevance that the system gives to the documents to the relevance that user would give (total error). Moreover, we can compute the partial errors taking into account on the one hand the relevant documents (relevant error) and the non relevant documents (non-relevant error).

\[
\text{Total error} = \frac{\sum_{i=1}^{\Omega} (u_i - s_i)^2}{\Omega} \\
\text{Relevant error} = \frac{\sum_{i=1}^{\Omega_r} (u_i - s_i)^2}{\Omega_r} \\
\text{Non relevant error} = \frac{\sum_{i=1}^{\Omega_{nr}} (u_i - s_i)^2}{\Omega_{nr}}
\]  

(9)

4.4. Results

As we can see in the tables below, the experiments has been realized with the two examples explained in section 4.2. In Table 1, the values of fuzzy recall-precision, fuzzy recall and fuzzy precision measures are shown using Zadeh’s and DSV’s cardinality. As can be observed this table, the difference between the values of the measures for both cardinalities can be noted, specially in the example 2, when the number of documents is higher (100). More exactly, the values of the measures for DSV’s cardinality are lower than the ones obtained with Zadeh’s cardinality. This fact would lead us to think that one of the measures is not as exact as we expect.

<table>
<thead>
<tr>
<th>Example</th>
<th>Cardinality</th>
<th>Fuzzy Recall - Precision</th>
<th>Fuzzy Recall</th>
<th>Fuzzy Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Zadeh</td>
<td>0.395</td>
<td>0.632</td>
<td>0.537</td>
</tr>
<tr>
<td></td>
<td>DSV</td>
<td>0.333</td>
<td>0.581</td>
<td>0.467</td>
</tr>
<tr>
<td>2</td>
<td>Zadeh</td>
<td>0.256</td>
<td>0.545</td>
<td>0.388</td>
</tr>
<tr>
<td></td>
<td>DSV</td>
<td>0.185</td>
<td>0.380</td>
<td>0.356</td>
</tr>
</tbody>
</table>

Table 1. Effectiveness fuzzy measures with different cardinalities

Let us observe Table 2, where the classification errors of the same examples with both cardinalities are calculated. In example 1, both errors are almost equal, although results with the DSV’s cardinality seem slightly higher. However, in example 2, the error for Zadeh’s cardinality is higher than DSV’s one, specially in the non relevant error.

This fact corroborates our assessment that Zadeh’s cardinality provides a high value for the cardinality of a fuzzy set due to an accumulation of small values.
5. Concluding Remarks and Future Work

In this work, we have presented new effectiveness information retrieval measures with fuzzy logic using alternatives to Zadeh’s cardinality. It has been shown that DSV cardinality provides more intuitive cardinalities with an efficiency of $O(n)$, with $n$ being the number of documents. Results have shown that the classification error is higher when the number of documents increase for measures with Zadeh’s cardinality, while the values for the fuzzy recall and fuzzy precision measures is higher than in measures with the DSV’s measures, which is contradictory.

In a future work, a deeply study of the experimental cases with different sets of documents with respect to their size and the number of relevant and non relevant documents within them will be presented.

<table>
<thead>
<tr>
<th>Example</th>
<th>Cardinality</th>
<th>Total error</th>
<th>Relevant error</th>
<th>Non Relevant Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Training</td>
<td>Test</td>
<td>Training</td>
</tr>
<tr>
<td>1</td>
<td>Zadeh</td>
<td>0.20</td>
<td>0.20</td>
<td>0.177</td>
</tr>
<tr>
<td></td>
<td>DSV</td>
<td>0.21</td>
<td>0.20</td>
<td>0.192</td>
</tr>
<tr>
<td>2</td>
<td>Zadeh</td>
<td>0.26</td>
<td>0.26</td>
<td>0.329</td>
</tr>
<tr>
<td></td>
<td>DSV</td>
<td>0.22</td>
<td>0.22</td>
<td>0.398</td>
</tr>
</tbody>
</table>

Table 2. Classification errors with different cardinalities.

References


<table>
<thead>
<tr>
<th>Page</th>
<th>Subject Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>123</td>
<td>India, 399, 403</td>
</tr>
<tr>
<td></td>
<td>Abstracts, 34, 79, 82, 133, 234, 238, 239, 482, 494, 498</td>
</tr>
<tr>
<td></td>
<td>Academic Libraries, 158, 159, 160, 163</td>
</tr>
<tr>
<td></td>
<td>Access to Information, 45, 80, 166, 213, 334, 411, 488, 506, 507</td>
</tr>
<tr>
<td></td>
<td>Alexandria Digital Library, 86, 232, 236</td>
</tr>
<tr>
<td></td>
<td>Algorithms, 46, 143, 251, 256, 301, 302, 359, 379, 422, 423, 426, 427, 428, 430, 431, 433, 434, 440, 446, 447, 452, 564, 572, 574, 577</td>
</tr>
<tr>
<td></td>
<td>Amazon.com, 279, 280, 281, 282, 283, 284</td>
</tr>
<tr>
<td></td>
<td>Amazonian Narratives, 546, 548</td>
</tr>
<tr>
<td></td>
<td>Analysis of actors, 412</td>
</tr>
<tr>
<td></td>
<td>Analysis of the Information, 74</td>
</tr>
<tr>
<td></td>
<td>Analytico-Synthetic Scheme, 113</td>
</tr>
<tr>
<td></td>
<td>Ananzi, 396, 399, 403</td>
</tr>
<tr>
<td></td>
<td>Application Programs, 338</td>
</tr>
<tr>
<td></td>
<td>Applications, 57, 66, 133, 137, 156, 234, 244, 303, 323, 343, 371, 385, 391, 443, 498, 571, 573, 585</td>
</tr>
<tr>
<td></td>
<td>Architecture Thesaurus, 47, 186, 191, 196, 197</td>
</tr>
<tr>
<td></td>
<td>Archive Finding Aids, 205</td>
</tr>
<tr>
<td></td>
<td>Archives, 89, 133, 136, 152, 205, 206, 208, 213, 214, 252, 415, 418</td>
</tr>
<tr>
<td></td>
<td>Archivists, 489, 492, 516</td>
</tr>
<tr>
<td></td>
<td>Artificial Intelligence Techniques, 305, 423</td>
</tr>
<tr>
<td></td>
<td>Artificial Neural Networks, 438, 573</td>
</tr>
<tr>
<td></td>
<td>Associative Structures, 305, 308, 309</td>
</tr>
<tr>
<td></td>
<td>Audiovisual Information, 120, 122, 125</td>
</tr>
<tr>
<td></td>
<td>Authorial Indexes, 358, 359, 360, 361</td>
</tr>
<tr>
<td></td>
<td>Authority Files, 57</td>
</tr>
<tr>
<td></td>
<td>Automatic Labeling, 205</td>
</tr>
<tr>
<td></td>
<td>Automatic Methods, 351, 357</td>
</tr>
<tr>
<td></td>
<td>Basic Level Concepts, 311, 312</td>
</tr>
<tr>
<td></td>
<td>Bayesian Belief, 443</td>
</tr>
<tr>
<td></td>
<td>Belief Networks, 443</td>
</tr>
<tr>
<td></td>
<td>Bibliographic Families, 107, 264, 267, 268, 531, 532, 536, 537</td>
</tr>
<tr>
<td></td>
<td>Bibliographic Heritage, 291</td>
</tr>
<tr>
<td></td>
<td>Bibliographic Relationship, 265</td>
</tr>
<tr>
<td></td>
<td>Browser, 88, 191, 192, 194, 381, 543</td>
</tr>
<tr>
<td></td>
<td>Buckley, 42, 43, 422, 425, 426, 428</td>
</tr>
<tr>
<td></td>
<td>Business Enterprises, 211, 212</td>
</tr>
<tr>
<td></td>
<td>Cadé, 396, 397, 399, 403</td>
</tr>
<tr>
<td></td>
<td>Cantor Set Theory, 181</td>
</tr>
<tr>
<td></td>
<td>Cataloging, 46, 79, 91, 173, 180, 196, 278, 507</td>
</tr>
<tr>
<td></td>
<td>Catalogues, 77, 296, 339</td>
</tr>
<tr>
<td></td>
<td>Categorisation, 112, 213, 214, 217, 386, 390, 391</td>
</tr>
<tr>
<td></td>
<td>CitEc Project, 251</td>
</tr>
<tr>
<td></td>
<td>Classification of Models, 91</td>
</tr>
<tr>
<td></td>
<td>Classification Research, 58, 59, 194</td>
</tr>
<tr>
<td></td>
<td>Classification Systems, 30, 43, 46, 47, 56, 103, 104, 128, 130, 136, 173, 318, 507</td>
</tr>
</tbody>
</table>
DOCUMENTALISTS, 516
DOMAIN ANALYSIS, 49, 56, 456
DTD - SGML, 205
ECOLOGICAL CLASSIFICATION SCHEMES, 412, 413, 417
ECONOMIC ACTIVITY, 50, 345, 349, 350
ECONOMIC CLASSIFICATION, 345, 346, 347, 348
ECONOMICS, 251, 252, 253, 482, 514
EFFECTIVE INFORMATION, 42, 404, 422
ELECTRONIC ENCODING, 205
ELECTRONIC ENVIRONMENTS, 28, 507
ELECTRONIC FORMATS, 109, 136, 251, 500
ELECTRONIC INFORMATION, 46, 136, 211, 440, 487, 523, 577
ELECTRONIC RESOURCES, 88, 137, 142, 145, 158, 160, 161, 162, 163, 164
ELECTRONIC RETRIEVAL, 28
ENCODED ARCHIVAL DESCRIPTION, 205
ENCYCLOPAEDISM, 457, 458, 463
ENCYCLOPAEDISTS, 457, 458, 459, 460
END USER, 155, 202, 493, 500
ENTERPRISES, 211
ENVIRONMENTAL SCIENCE, 332, 335
ETHICAL VALUES, 487, 488, 489, 491
ETHICS, 54, 85, 463, 487, 490, 491, 492, 518, 519, 520, 521
EVALUATION METHODS, 386
EXAMPLE TECHNIQUE, 429
EXOMEMORY, 516, 517, 518, 519, 520
EXTENDED BOOLEAN QUERIES, 429
FACET ANALYSIS, 77, 86, 137, 138, 139, 141, 194, 196, 336
FACETED SYSTEMS, 191
FACETED THESAURS, 191
FAITHS, 539
FILM COLLABORATORY, 412
FILM INDEXING, 412, 415, 416, 417, 421
FILTERING INFORMATION, 358, 572
FILTERING METHODOLOGY, 552
FILTERING TECHNIQUES, 552
FINDING AIDS, 205, 207, 209, 213
FORMAL STRUCTURES, 308
FRENCH-LANGUAGE, 271
FUZZY LINGUISTIC, 372, 373, 377, 378, 568, 569, 570, 571
GCL TERMS, 142
GENDER STUDIES, 393, 394, 399
GENETIC ALGORITHMS, 422
GLOBAL ECONOMY, 345
GLOBAL INFORMATION SYSTEMS, 303, 341, 343, 344
GLOSSARY OF TERMS, 539, 544
GRAPHICAL MODELS, 443, 444
GRAPHICAL TABLE OF CONTENTS, 437
GRAPHICS, 120, 219, 282
HEALTH RESOURCES, 198
HERMES, 386, 388, 389
HETEROGENEOUS DATABASES, 304
HOLISTIC APPROACH, 74
HYBRID SEARCH ENGINES, 245
HYPERTEXTUAL THESAURIS, 493
IBERMARC, 264, 268
IDE DEC-HI METHOD, 422, 423, 425, 426, 427
IFLA, 130, 165, 170, 171, 172, 261, 263, 264, 267, 269, 278, 317, 392
IMAGE ANALYSIS, 120
INDEXING LANGUAGES, 214, 409
INDUCTIVE QUERY, 427, 429, 430, 435
INFORMATION ACCESS, 47, 51, 149, 150, 151, 153, 173, 404, 571, 575
INFORMATION AGENTS, 379, 380, 381, 382, 384, 386
INFORMATION FILTERING, 151, 386
INFORMATION PROCESS OF FICTION, 74, 75
INFORMATION PROCESSING, 57, 74, 121, 155, 230, 235, 305
INFORMATION RECOVERY SYSTEM, 286, 287
MAPS CATEGORIES, 191
MASIR, 379, 382, 383, 384
MEANS-ENDS MODEL, 412, 413, 416
MERCOSUR, 393
METADATA, 47, 57, 86, 90, 100, 103, 104, 105, 128, 130, 133, 135, 140, 141, 142, 144, 145, 196, 198, 200, 203, 214, 217, 222, 223, 237, 251, 253, 255, 298, 301, 305, 360, 361, 507, 510, 524
METAFILTER, 552, 557, 558
METAPHORICAL REPRESENTATIONS, 553, 555, 556, 557
META-THEAURUS, 143, 144, 147
MICROSOFT TECHNOLOGY, 205
MONOGRAPHS, 170, 173
MULTICULTURALISM, 333, 516, 519
MULTIDISCIPLINARY KNOWLEDGE, 54
MULTI-GRANULAR LINGUISTIC, 372, 373, 374, 375, 376, 377
MULTIHIERARCHICAL STRUCTURE, 279
MULTI-TERM QUERIES, 191
MULTILINGUAL INFORMATION ACCESS, 387, 388, 404
MULTILINGUAL SYSTEM, 331, 386
MULTI-TERM QUERIES, 191
MULTIWORD RECOGNITION, 559
NARRATIVE DOCUMENTS, 49, 93
NATURAL LANGUAGE, 15, 16, 18, 20, 24, 76, 93, 114, 121, 122, 148, 149, 152, 153, 154, 200, 230, 305, 307, 309, 315, 339, 404, 405, 410, 411, 501, 504, 548, 571
NATURAL LANGUAGE PROCESSING, 26, 99, 114, 156, 559, 564
NON-COOPERATIVE AGENTS, 379
ONLINE ACCESS TOOLS, 54
ONLINE INFORMATION, 158, 159
ONTOLOGICAL ANALYSIS, 67
ONTOLOGICAL CATEGORIES, 225, 226, 228, 230
ONTOLOGICAL CLASSIFICATIONS, 225, 226
OPACS, 108, 437
ORAL TRADITION, 546, 547
ORGANIZATIONAL SCHEMES, 158, 160, 161, 163, 164
PARADIGMS, 465, 466, 471
PATRONY, 291, 294, 296
PERSONALISATION, 386, 390, 523, 525, 528, 529
PRACTICAL METHODS, 205
PRACTICAL MODELS, 111
PROFESSIONAL ETHICS, 13
PUBLIC SECTOR, 142
QUALITATIVE ANALYSIS, 269, 271, 328, 393, 397, 532, 537, 538, 546
QUANTITATIVE ANALYSIS, 498, 532, 543, 544
QUERY BUILDER, 191
QUERY EXPANSION, 151
QUERYING, 93
RANGANATHAN, 117, 118, 119, 137, 138, 139, 141, 194, 319, 323, 450, 451, 472, 481, 552, 553, 555, 556, 557, 558
REBIUN, 264, 269, 296, 502
RECORDED KNOWLEDGE, 81, 265, 332, 530
RELEVANCE FEEDBACK, 422, 423
REPRESENTATION MODELS, 240
RESIDUAL COLLECTION METHOD, 422, 426
RETRIEVAL NEEDS, 479, 481
RETRIEVAL STATUS VALUES, 372
SCHOLARLY COMMUNICATION, 28, 29
SCHOLARLY DOCUMENTS, 28
SCIENTIFIC DOMAIN, 80
SCIENTIFIC INFORMATION, 149, 203, 571
SCIENTIFIC MODELS, 86, 88, 89, 90, 91
SEARCHERS, 36, 130, 144, 148, 196, 245, 279, 430, 500
SEMANTIC CLASSIFICATIONS, 225, 226
SEMANTIC UNIVERSALS, 17, 18, 19, 20, 26
SEMIOTIC CLASSIFICATIONS, 225
SETS THEORY, 584
SIMULATED-ANNEALING, 429
SOCIAL GROUPS, 102, 552, 553, 554, 557
SOFTWARE DEVELOPMENT AND REUSE, 218
SOFTWARE SYSTEM, 251

SOUTH AFRICA, 159, 164, 211, 212, 217, 396, 397, 403
SPANISH UNIVERSITIES, 269, 291
SPECIAL CLASSIFICATIONS, 194, 450, 485
SPECIALISED LITERATURE, 443
STATISTICAL DATA, 318, 320, 348, 532
STATISTICS, 74, 302, 320, 322, 438, 440, 512
SUBJECT ACCESS, 54, 55, 136, 173, 278, 317, 500, 501, 507
SUBJECT CATALOGUING, 165, 167
SUBJECT HEADING SYSTEMS, 127
SUBJECT MAPS, 437
SUBJECT ONTOGENY, 54, 55, 56, 57, 58
SUBJECT PORTAL, 135
SYMBOLIC ONTOGENY, 54, 55, 56, 57, 58
TAYLOR, 37, 43, 192, 197, 213, 217, 457, 464, 533, 536, 574, 577
TECHNOLOGICAL MEDIATIONS, 50
TECHNOLOGICAL MEDIATIONS, 50
TERMINOLOGIES, 137, 152, 393, 394
THEMATIC IDENTITIES, 553
THESAURAL KNOWLEDGE STRUCTURES, 493
List of Contributors

Alba Quiñones, Francisco de
C/ Juan Antonio de Vizarrón s/n, 41092 Sevilla, Spain
Phone: 955065299-955065139
Fax: 955065446
frankaq@hotmail.com

Albrechtsen, Hanne
Centre for Human-Computer Interaction, Risø National Laboratory Denmark
P.O.Box 49, DK-4000 Roskilde Denmark
Phone: +45 46 77 51 53
Fax: +45 46 77 51 99
hanne.albrechtsen@risoe.dk

Aldana, J,F.
Departamento de Lenguajes y Ciencias de la Computación. Universidad de Málaga
Campus de Teatinos, 29071 Málaga, Spain
Phone: +34 (9)582813
Fax: 952131397
jfam@lcc.uma.es

Andersen, Jack
Department of Information Studies, Royal School of Library and Information Science
Birketinget 6, 2300 Copenhagen S, Kontor: C. 418, Denmark
Phone: 32 58 70 77 lokal 493
jan@db.dk

Angôs Ullate, José Mª
Centro de Documentación Científica, Universidad de Zaragoza
C/ Pedro Cerbuna, 12, Zaragoza – 50009, Spain
Phone: +34(9)76761332
Fax: +34(9)76761088
angos@posta.unizar.es

Asghar Shiri, Ali
Department of Computer and Information Sciences, University of Strathclyde
26 Richmond Street, Livingstone Tower, Glasgow G1 1XH UK
Phone: 141- 548 4065 or 141-548 3700
Fax: 141- 553 1393
shiri@dis.strath.ac.uk

Assadi, Houssem
France Télécom
R&D – DIH/UCE,38-40, rue du Général Leclerc, 92 794 Issy-les-Moulineaux
Cedex 9, France
houssem.assadi@rd.francetelecom.com

Banwell, Linda M.
Information Management Research Institute, School of Information Studies, Northumbria University
Newcastle upon Tyne NE1 8ST, England
Phone: 0191-227-3768
Fax: 0191-227-3671
Linda.Banwell@northumbria.ac.uk

Barité Roqueta, Mario Guido
Escuela de Biblioteconomía. Univ. de la República Oriental del Uruguay
Edificio Central: Emilio Frugoni 1427, 11.200 Montevideo, Uruguay
Phone: (598 2) 401 07 88
Fax: (598 2) 400 58 10
baritema@eubca.edu.uy

Barrueco, José Manuel
Universitat de València
Spain
Biblioteca de Ciencias Sociales, Universitat de València, Valencia 46010, Spain
Phone: +34 -963733971
Jose.Barrueco@uv.es

Bean, Carol A.
National Library of Medicine
6705 Rockledge Drive, Rockledge 1, Suite 301, MSC 7968, Bethesda, MD 20892, USA
Phone: +1-301-496-3113
Fax: +1-301-496-402-2952
bean@nlm.nih.gov

Beauvisage, Thomas
France Télécom
R&D – DIH/UCE,38-40, rue du Général Leclerc, 92 794 Issy-les-Moulineaux
Cedex 9, France
Phone/Fax: 01 43 56 94 03
thomas.beauvisage@rd.francetelecom.com

Beghtol, Clare
Faculty of Information Studies, University of Toronto
140 St. George St.  
Toronto  
Ontario, Ontario M5S 3G6 Canada  
Phone: (416) 978-8852  
Fax: (416) 971-1399  
beghtol@fis.utoronto.ca

Binding, Ceri  
Hypermedia Research Unit, University of Glamorgan  
Pontypridd, CF37 1DL, Wales, UK  
Phone: +1443-482271  
Fax: +1443-482715  
cbinding@glam.ac.uk

Blanco Medina, Ignacio José  
Departamento de Ciencias de la Computación e Inteligencia Artificial. Universidad de Granada.  
C/Daniel Saucedo s/n, 18071, Granada, Spain  
Phone: +34 958244019  
Fax: +34 958243317  
iblanco@decsai.ugr.es

Blocks, Dorothee  
Hypermedia Research Unit, University of Glamorgan  
Pontypridd, CF37 1DL, Wales, UK  
Phone: +1443-482271  
Fax: +1443-482715  
dblocks@glam.ac.uk

Broughton, Vanda  
School of Library, Archive & Information Studies, University College London  
Gower Street, LONDON, WC1E 6BT, United Kingdom  
Phone: 020 7679 2291  
v.broughton@ucl.ac.uk

Campbell, Grant  
Faculty of Information and Media Studies, University of Western Ontario  
London, Ontario, Canada N6A 5B7  
Fax: (519) 661-3506  
gcampbell@uwo.ca

Campos Ibáñez, Luis Miguel de  
Departamento de Ciencias de la Computación e Inteligencia Artificial. Universidad de Granada.
Ourense
32004 - As Lagoas - Ourense Spain
Phone:+34 988 387018
Fax: +34 988 387001
pcuesta@uvigo.es

Cunliffe, Daniel
Hypermedia Research Unit, University of Glamorgan
Pontypridd,CF37 1DL, Wales, UK
Phone: +1443-482271
Fax: +1443-482715
djcunlif@glam.ac.uk

Chaudiron, Stéphane
Département Information-Communication (Ilphi), Université de Paris12
200 avenue de la république, 92001 Nanterre Cedex
Phone: (33) - 01 40 97 73 54
Fax: (33) - 01 40 97 59 81
stephane.chaudiron@u-paris10.fr

Chowdhury, Gobinda
Department of Computer and Information Sciences, University of Strathclyde
26 Richmond Street, Livingstone Tower, Glasgow G1 1XH UK
Phone: 141- 548 3092
Fax: 141- 553 1393
gobinda@dis.strath.ac.uk

Dextre Clarke, Stella G.
Information consultant
West Hendred, Wantage, Oxon, OX12
8RR United Kingdom
Phone: 01235 833 298
Fax: 01235 863 298
SDClarke@LukeHouse.demon.co.uk

Díaz Esteban, Alberto
CEES Felipe II
Despacho 37 - Edificio C
28670 - Villaviciosa de Odón, Aranjuez Madrid. Spain.
Phone: +34 916169400 Ext. 671
alberto@dinar.esi.uem.es

Díaz Galiano, M. C.
Departamento. de Informática, Universidad de Jaén. Avda. Madrid, 35, 23071, Jaén (Spain)
Phone: +34 953 012382

Díaz de Oliveira, Maria Odaisa
Universidade Federal do Pará/Brasil
Av. Governador José Malcher, 2271, apto. 602, 66.060-230, Belém – Pará - Brasil
Phone: 91 246-74-78
Fax: 91 211-13-55
odaisa@ufpa.br

Dippie, Shona R.
School of Library & Information Studies, University of Alberta
3-20 Rutherford South, University of Alberta, Edmonton, Alberta, Canada T6G 2J4
Phone: 780 492-4578

Elto Brun, Ricardo
Departamento de Biblioteconomía y Documentación. Universidad Carlos III de Madrid.
C/ Madrid 126-128. 28903 Getafe Madrid. Spain
Phone: +34 91 6249239
Fax: +34 91 6249212
reito@bib.uc3m.es

Elchirigoity, Fernando
Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign
501 E. Daniel St., Champaign, IL 61820-6211
Phone: 217-244-7365
Fax: 217-244-3302
elichi@uiuc.edu

Espinheiro de Oliveira, Maria Odaisa
Universidade Federal do Pará/Brasil
Av. Governador José Malcher, 2271, apto. 602, 66.060-230, Belém – Pará - Brasil
Phone: 91 246-74-78
Fax: 91 211-13-55
odaisa@ufpa.br

Faba Pérez, Cristina
Facultad de Biblioteconomía y Documentación. Universidad de Extremadura, Alcazaba de Badajoz s/n, 06071 Badajoz, Spain
cfabper@alcazaba.unex.es

Fattahi, Rahmatollah
Department of Library and Information Studies, Faculty of Education and Psychology Ferdowsi University of Mashhad
Mashhad, Iran, P. O. Box 1518
Phone: (9851) 812008; 832237
Fax: (9851) 827079 or (9851) 812008
fattahi@ferdowsi.um.ac.ir

Fernández Falero, Mª del Rosario
University of Extremadura, Faculty of Librarianship Studies and Documentation
La Alcazaba de Badajoz, S/N, 06071-Badajoz (Spain)
Phone: 924 289300 (ext. 6426)
Fax: 924 286401
rferfal@alcazaba.unex.es

Fernández-Luna, Juan M.
Departamento de Informática.
Escuela Politécnica Superior. Universidad de Jaén. 23071 – Jaén, Spain
Phone: +34 953026559
jmfluna@ujaen.es

Fernández-Molina, J. Carlos
Universidad de Granada
Spain.
Colegio Máximo de Cartuja. Campus Universitario de Cartuja s/n. 18071 - Granada
Phone: +34 958243940
jcfernana@ugr.es

Fernández Ruiz, Mª Jesús
Sección Información Web, Ayuntamiento de Zaragoza
C/ Plaza de las Catedrales 18, Zaragoza - 50071 (España)
Phone: +34 976721278
Fax: +34 976721166
mjferruiz@ayto-zaragoza.es

Frâncu, Victoria
Central University Library 1, Boteanu Street, Bucharest, Romania
Phone: 0040-1-312.06.61
Fax: 0040-1-312-01-08
francu@bcub.ro

Frias, José Antonio
Departamento de Biblioteconomía y Documentación, Universidad de Salamanca
Francisco Vitoria, 6-16
37008 Salamanca. Spain
Phone: 34-923-294-580
Fax: 34-923-294-582
frias@gugu.usal.es

Furner, Jonathan
Department of Information Studies, University of California, Los Angeles, USA
GSE&Building, Box 951520, Los Angeles, California 90095-1520, USA
Phone: (310) 825-5210
jfurner@ucla.edu

Fusco, Lisa M.
University of Washington, Information School, Seattle
Box 352840, Seattle, WA 98195-2840, USA
Phone: (206) 523-9391
Fax: (206) 616-3152
lisaf59@hotmail.com

García Cabrera, Lina
Departamento de informática, Escuela Politécnica Superior, Universidad de Jaén. Spain
Phone: +34 953 002475
lina@ujaen.es

García Gutiérrez, Antonio
Departamento de Periodismo. Facultad de Ciencias de la Información. Universidad de Sevilla. C/ Gonzalo de Bilbao, 7 y 9 41003 Sevilla. Spain
Phone: +34 954 48 60 57
algarcia@cica.es

García Jiménez, Antonio
Universidad Rey Juan Carlos
Madrid. Spain
Phone: +34 91 486 33 83
antonio.garcia@cccom.urjc.es

Gervás, Pablo
Facultad de CC. Matemáticas. Universidad Complutense de Madrid
C/ Avenida de la Complutense s/n 28040 Madrid
Phone: +34 913944293
pgervas@sip.ucm.es

Gjerulf Eslau, Anna
H. Lundbeck
H. Lundbeck A/S, Ottliavej 7-9, DK-2500 Valby
Phone: +45 36 30 13 11
<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation</th>
<th>Address</th>
<th>Phone</th>
<th>Fax</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hudon, Michèle</td>
<td>École de bibliothéconomie et des sciences de l'information, Université de Montréal</td>
<td>CP 6128, succ. Centre-Ville, Montréal, QC H3C 3J7, CANADA</td>
<td>Phone: (+1-514)-343-6046</td>
<td>Fax: (+1-514)-343-5753</td>
<td><a href="mailto:michelle.hudon@umontreal.ca">michelle.hudon@umontreal.ca</a></td>
</tr>
<tr>
<td>Huete Guadix, Juan F.</td>
<td>Departamento de Ciencias de la Computación e Inteligencia Artificial. Universidad de Granada.</td>
<td>C/Daniel Saucedo s/n, 18071, Granada, Spain</td>
<td>Phone: (+34)958244019</td>
<td>Fax: (+34)958243317</td>
<td><a href="mailto:jhg@decsai.ugr.es">jhg@decsai.ugr.es</a></td>
</tr>
<tr>
<td>Hurtado Torres, Mª Visitation</td>
<td>Departamento de Lenguajes y Sistemas Informáticos.ETS de Ingeniería. Informática Universidad de Granada. C/ Periodista Daniel Saucedo Aranda s/n 18071. Granada. Spain</td>
<td>Phone: (+34)958243697</td>
<td><a href="mailto:mhurtado@ugr.es">mhurtado@ugr.es</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ihadhjadeden, Majid</td>
<td>Département Information-Communication (lphi), Université de Paris11</td>
<td>200 avenue de la république, 92001 Nanterre Cedex</td>
<td>Phone: (33) - 01 40 97 73 54</td>
<td>Fax: (33) - 01 40 97 59 81</td>
<td><a href="mailto:ihadjade@u-paris10.fr">ihadjade@u-paris10.fr</a></td>
</tr>
<tr>
<td>Inglada, Vicente Julián</td>
<td>Departamento de Sistemas Informáticos y Computación. Universidad Politécnica de Valencia. Camino de Vera s/n</td>
<td>P.O. Box: 22012 E-46020 - Valencia Spain</td>
<td>Phone: (+34)963877350</td>
<td>Fax: (+34)963877359</td>
<td><a href="mailto:vinglada@dsic.upv.es">vinglada@dsic.upv.es</a></td>
</tr>
<tr>
<td>Iyer, Hemalata</td>
<td>School of Information Science and Policy, University at Albany</td>
<td>SUNY, Albany, NY 12222, USA</td>
<td>Phone: (518) 442-5116</td>
<td><a href="mailto:Hi651@csc.albany.edu">Hi651@csc.albany.edu</a></td>
<td></td>
</tr>
<tr>
<td>Jacob, Elin K.</td>
<td>Indiana University-Bloomington</td>
<td>1320 East 10th, Main Library 011, Bloomington, IN 47405-3907, USA</td>
<td>Phone: 812-855-4671</td>
<td>Fax: 812-855-6166</td>
<td><a href="mailto:ejacob@indiana.edu">ejacob@indiana.edu</a></td>
</tr>
<tr>
<td>Kedar, Rochelle</td>
<td>Department of Information Science, Bar-Ilan University</td>
<td>Ramat-Gan 52900,Israel</td>
<td>Phone: +972-9-7416096</td>
<td>Fax: +972-9-7480181</td>
<td><a href="mailto:kedarr@mail.biu.ac.il">kedarr@mail.biu.ac.il</a></td>
</tr>
<tr>
<td>Keefe, Jeanne M.</td>
<td>Visual Resources Librarian, Architecture Library, Rensselaer Polytechnic Institute</td>
<td>Troy, NY 12180</td>
<td>Phone: (518) 276-2727</td>
<td><a href="mailto:keefej@rpi.edu">keefej@rpi.edu</a></td>
<td></td>
</tr>
<tr>
<td>Khammaci, T.</td>
<td>IRIN, Université de Nantes</td>
<td>4 rue de la Houssinière, BP 92208 , 44322 Nantes Cedex 03, France</td>
<td>Phone: (33-2) 51-12-58-34</td>
<td>Fax: (33-2) 51-12-58-12</td>
<td><a href="mailto:Tahar.Khammaci@irin.univ-nantes.fr">Tahar.Khammaci@irin.univ-nantes.fr</a></td>
</tr>
<tr>
<td>Khoo, Christopher S.G.</td>
<td>Division of Information Studies, School of Communication &amp; Information, Nanyang Technological University</td>
<td>31 Nanyang Link, Singapore 637718</td>
<td>Phone: (65) 6790 4602</td>
<td>Fax: (65) 6791-5214</td>
<td><a href="mailto:assgkhoo@ntu.edu.sg">assgkhoo@ntu.edu.sg</a></td>
</tr>
<tr>
<td>King, Christopher</td>
<td>Department of Computing and Electrical Engineering, Heriot-Watt University, Edinburgh</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knott Malone, Cheryl</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Kobashi, Nair Yumiko
Universidade de São Paulo/Escola de
Comunicações e Artes/Departamento de
Biblioteconomia e Documentação
Av. Prof. Lúcio Martins Rodrigues 443,
Cidades Universitária 05508-900, São
Paulo, SP, Brazil
Phone: 55-11-3091-4076
Fax: 55-11-3091-4325
nykobash@usp.br

Kwasnik, Barbara H.
School of Information Studies, Syracuse
University
4-283 Center for Science and Technology,
Syracuse, NY 13244-4100, USA
Phone: 315-443-4547
Fax: 315-443-5806
bkwasnik@syr.edu

Lee, Hur-Li
University of Wisconsin-Milwaukee
Milwaukee, Wisconsin, USA
Bolton Hall, Room 526
Phone: (414) 229-6838
Fax: (414) 229-4848
hurli@uwm.edu

Lei Zeng, Marcia
School of Library and Information
Science, Kent State University
Kent, OH 44242-0001, U.S.A.
Phone: (330) 672-0009 or -2782
Fax: (330) 672-7965
mzeng@kent.edu, (phone) (fax)

Lykke Nielsen, Marianne
Royal School of Library and Information
Science
Birketingsgade 6, DK-2300, København S.
Phone: +45 32 58 60 66
Fax: +45 32 84 02 01
mln@db.dk

López-Huertas, María José
Facultad de Biblioteconomía y
Documentación. Univ. de Granada
Colegio Máximo de Cartuja. Campus
Universitario de Cartuja s/n. 18071 –
Granada. Spain.
Phone: 958-243936
mjlopez@ugr.es

López-Pujalte, Cristina
Library and Information Science Faculty,
University of Extremadura, Alcazaba de
Badajoz s/n, 06071 Badajoz, Spain
clopez@alcazaba.unex.es

Madalli, Devika P.
Documentation Research and Training
Centre, Indian Statistical Institute
8th Mile, Mysore Road, Bangalore -- 560
059, Karnataka, India
Phone: 91-80-8483002/3/4/6 Extn: 497
Fax: 91-80-8484265 OR 8483001
devika@isibang.ac.in

Mai, Jens-Erik
The Information School. University of
Washington
Mary Gates Hall
Box 352840
Seattle, WA 98195-2840
Phone: (206) 616-2541
Fax: (206) 616-3152
jemai@u.washington.edu

Mark Pejtersen, Annelise
Centre for Human-Computer Interaction,
Risø National Laboratory Denmark
P.O.Box 49, DK-4000 Roskilde Denmark
Phone: +45 46 77 51 49
Fax: +45 46 77 51 99
annelise.mark.pejtersen@risoe.dk

McIlwaine, I.C.
University College London
Gower Street, London WC1E 6BT
Phone: 0171-387 7050; 0171-380 7204
Fax: +44 171 383 0557
o.manager@ucl.ac.uk

Martín-Bautista, M.J.
Departamento de Ciencias de la
Computación e Inteligencia Artificial.
Universidad de Granada.
C/Daniel Saucedo s/n, 18071, Granada, Spain
Phone: +34 958244019
Fax: +34 958243317
mbautista@decsai.ugr.es

Martín Valdivia, M.T.
Departamento de Informática, Universidad de Jaén, Avda. Madrid, 35, 23071, Jaén (Spain)
Phone: +34 953 01.22.02
maite@ujaen.es

Martínez, F.
Departamento de Informática, Universidad de Jaén, Avda. Madrid, 35, 23071, Jaén (Spain)
Phone: +34 953 00 24 77
Fax: +34 953 00 24 20

Marwick, David H.
Department of Computing and Electrical Engineering, Heriot-Watt University, Edinburgh

Moneda Corrochano, Mercedes de la Universidad de Granada
Colegio Máximo de Cartuja. Campus Universitario de Cartuja s/n. 18071 - Granada. Spain
Phone: +34 958243940
dlmoneda@ugr.es

Moreno, N.
Department of Lenguajes y Ciencias de la Computación. University of Málaga, Campus de Teatinos 29071 Málaga. Spain
Phone: 952132865
vergara@lcc.uma.es

Moros Ramírez, Ángel
Universidad Central de Venezuela.
angelmoros@hotmail.com

Moya-Anegón, Félix
Departamento de Biblioteconomía y Documentación Universidad de Granada. Colegio Máximo de Cartuja. Campus Universitario de Cartuja s/n. 18071 - Granada. Spain
Phone: +34 958243940
felix@ugr.es

Muñoz Fernández, Francisco José
Centro de Documentación Europea. Universidad de Granada
Rector López Argüeta s/n
18071 - Granada. Spain
Phone:+34 (9) 58248351
Fax:+34 (9) 58242382
franjomi@hotmail.com

Mustafa el Hadi, Widad
UFR IDIST & CERSATES (CNRS UMR 8529), Université Charles De Gaulle, Lille 3
BP 149, F-59 653 Villeneuve D'Ascq, France
Phone: (33) 3 20 41 68 15
Fax: (33) 3 20 41 63 79
mustafa@univ-lille3.fr

Naumis Peña, Catalina
Centro Universitario de Investigaciones Biblioteconómicas, Universidad Nacional Autónoma de México
Piso 12, Torre II de Humanidades. Cd. Universitaria, CP 04510 Coyoacán, México D. F.
naumis@servidor.unam.mx

Nebro, A.J.
Department of Lenguajes y Ciencias de la Computación. University of Málaga
Campus de Teatinos 29071 Málaga. Spain
Phone: +34 95233310
Fax: +34 952131397
antonio@lcc.uma.es

Neelameghan, A.
216, 4th Main Road, 16th Cross Road, Bangalore 560055, India
anm@satyam.net.in

Negrini, Giliola
Consiglio Nazionale delle Ricerche - ISRDS, Roma, Italy
Via Cesare De Lollis 12
00185 - Roma , Italy
Phone: +39 06.448791
Fax: +39 06.4463836
negrini@isrds.rm.cnr.it

Ng, Karen
Nielsen, Juliet
School of Library & Information Studies, University of Alberta
Canada
jnielsen@ualberta.ca

Olson, Hope A.
School of Library & Information Studies, University of Alberta
3-20 Rutherford South, University of Alberta, Edmonton, Alberta, Canada T6G 2J4
Phone: 780 492-4578
Fax: 789 492-2430
slis@ualberta.ca

Ou, Shlyan
Division of Information Studies, School of Communication & Information, Nanyang Technological University
31 Nanyang Link, Singapore 637718
Phone: (65) 6790 4602
Fax: (65) 6791-5214
pg00096125@ntu.edu.sg

Oussalah, M.
IRIN, Université de Nantes
2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex 03, France
Phone: (33-2) 51-12-58-47
Fax: (33-2) 51-12-58-12
Mourad. Oussalah@irin.univ-nantes.fr

Parets Llorca, José
Departamento de Lenguajes y Sistemas Informáticos, ETS Ingeniería Informática, Universidad de Granada
C/Periodista Daniel Saucedo Aranda s/n, 18071 Granada. Spain
Phone:+34 958240574
jparets@ugr.es

Parirokh, Mehri
Department of Library and Information Studies, Ferdowsi University of Mashhad

Mashhad, Iran
parirokh@ferdowsi.um.ac.ir

Peis, Eduardo
Facultad de Biblioteconomía y Documentación. Universidad de Granada. Colegio Máximo de Cartuja. Campus Universitario de Cartuja s/n. 18071 Granada. Spain
Phone: +34 958243940
epeis@ugr.es

Pérez López, Ana
Facultad de Biblioteconomía y Documentación. Universidad de Granada. Colegio Máximo de Cartuja. Campus Universitario de Cartuja s/n. 18071- Granada. Spain
Phone: +34 958243940
aperez@ugr.es

Piero Zarri, Gian
Centre National Recherche Scientifique EHESS-CAMS
54. Boulevard Raspail
F-75270 Paris
France
Phone: +33-1-49.54.20.33
Fax: +33-1-49.54.21.09
Email: zarri@cams.msh-paris-fr

Polanco, Xavier
Unité de Recherche et Innovation, INIST-CNRS
2, Allée du Parc de Brabois
F-54514 Vandoeuvre-lès-Nancy
Phone: +33 (0)3 83 50 46 00
Fax:+33 (0)3 83 50 46 50
polanco @inist.fr

Poli, Roberto
Department of Sociology and Social Research, Trento and Mitteleuropa Foundation
26, Verdi street
38100 Trento, Italy
Phone: ++39-461-881-403
Fax: ++39-461-881-348
poli@riscl.gelso.unitn.it

Prasad, A.R.D.
Documentation Research and Training Centre, Indian Statistical Institute
8th Mile, Mysore Road, Bangalore -- 560 059, Karnataka, India
Phone: 91-80-8483002/3/4/6 Extsn: 496
Fax: 91-80-8484265 OR 8483001
ard@isibang.ac.in

Priss, Uta
School of Library and Information Science, Indiana University
Bloomington, USA
Phone: (812) 855-2793
upriss@indiana.edu

Quintero Orta, Ana
Biblioteca Universitaria. Universidad de Huelva. Huelva. Spain
aquintero@onuba.com

Qin, Jian
School of Information Studies, Syracuse University
4-232 CST
Syracuse, NY 13244, USA
Phone: (315)443-5642
Fax: (315)443-5806
jqin@syr.edu

Reyes Barragán, María J.
Library and Information Science Faculty, University of Extremadura, Alcazaba de Badajoz s/n, 06071 Badajoz, Spain
mjreyes@alcazaba.unex.es

Riesthuis, Gerhard J.A.
Leerstoelgroep BAI
Universiteit van Amsterdam
Oude Turfmarkt 147
1012 GC Amsterdam
Phone: + 31 (0)20 525 22 95
Fax: + 31 (0)20 525 20 97

Ríos Hilarío, Ana Belén
Departamento de Biblioteconomía y Documentación. Universidad de Salamanca
Spain
anarihi@gugu.usal.es

Revie, Crawford
Department of Computer and Information Sciences, University of Strathclyde
26 Richmond Street, Livingstone Tower, Glasgow G1 1XH UK
Phone: 141- 548 3301
Fax: 141- 553 1393
crawford@dis.strath.ac.uk

Rocha, João
Departamento de Engenharia Informática, Instituto Superior de Engenharia do Porto, Instituto Politécnico do Porto
Rua Dr. António Bernardino de Almeida, 431, 4200 - 072 Porto, PORTUGAL
Phone: +351 228340500
Fax: +351 228321159
Joao.Rocha@dei.isep.ipp.pt

Rivas Santos, V. M.
Departamento de Informática, Universidad de Jaén. Avda. Madrid, 35, 23071, Jaén. Spain
Phone: +34 953 012344
Fax: +34 953 00 24 20
vrivas@ujaen.es

Rodríguez, Francisco J.
Departamento de Informática
Escuela Superior de Ingeniería Informática
Edificio Politécnico
Universidade de Vigo - Campus de Ourense
32004 - As Lagoas -Ourense.Spain
Phone: +34 986 812411
franjrm@uvigo.es

Roldán, M.M.
Department of Lenguajes y Ciencias de la Computación. University of Málaga
Campus de Teatinos
29071 Málaga. Spain
Phone: 952132865
Fax: 952131397
mmar@lcc.uma.es

Role, François
Département Information-Communication (lphi), Université de Paris 10
200 avenue de la république, 92001 Nanterre Cedex
Phone: (33) - 01 40 97 73 54
Fax: (33) - 01 40 97 59 81
françois.role@recherche.gouv.fr

Ruiz, Antonio
Facultad de Biblioteconomía y Documentación. Universidad de Granada.
Colegio Máximo de Cartuja, Campus Universitario de Cartuja s/n, 18071 – Granada. Spain
Phone: +34(9)58243940
aruizr@ugr.es

Saarí, Jarmo
Development Manager
Kuopio University Library
P.O. BOX 1627, FIN-70211 Kuopio, Finland
Phone: +358-17-163475, +358-50-3637904
Fax: +358-17-163410
jarmo.saarit@uku.fi

Saaví, José Antonio
Departamento de Ciencias de la Documentación e Historia de la Ciencia, Universidad de Zaragoza
C/ Pedro Cerbuna, 12, Zaragoza - 50009 (España)
Phone: 976761000 ext. 3562
Fax: 976762305
jaso@posta.unizar.es

Sánchez Fernández, Daniel
Departamento de Ciencias de la Computación e Inteligencia Artificial. Universidad de Granada.
C/Daniel Saucedo s/n, 18071, Granada, Spain
Phone: +34 958244019
Fax: +34 958243317
daniel@decsai.ugr.es

Sánchez Domínguez, María del Carmen
Facultad de Documentación. Universidad de Alcalá de Henares.
Aulario María de Guzmán. C/ San Cirilo, s/n. 28201 Alcalá de Henares (Madrid)
Phone: 91-885.50.04
Fax: 91885.50.05
mc.sdomin@uah.es

Serrano Cobos, Jorge
Planeta DeAgostini On-Line
Av. Pau Casals 14, 5ª planta 08021 Barcelona
Phone: +34 93-2405907
jserrano@temalia.com

Shapiro, Jeremy J.
Fielding Graduate Institute
302 West 86th Street, #3B, New York, NY 10024-3146, U.S.A
Phone: (212) 595-8482
jshapiro@fielding.edu

Shoham, Snunith
Department of Information Science, Bar-Ilan University
Ramat-Gan 52900, Israel
Phone: +972-8-9721739; +972-54-965394
Fax: +972-8-9721518
shohas@mail.biu.ac.il

Silva, Nuno
Departamento de Engenharia Informática, Instituto Superior de Engenharia do Porto, Instituto Politécnico do Porto
Rua Dr. António Bernardino de Almeida, 431, 4200 - 072 Porto, PORTUGAL
Phone: +351 228340500
Fax: +351 228321159
Nuno.Silva@dei.isep.ipp.pt

Slavic, Aida
SLAIS, University College London, UK
Gower Street, London, WC1E 6BT, UK
aida.slavic@ucl.ac.uk

Smiraglia, Richard P.
Palmer School of Library and Information Science, Long Island University
720 Northern Blvd., Brookville NY 11548 USA
Phone: (516) 299-2174
Fax: (516) 299-4168
Richard.Smiraglia@liu.edu

Smith, Terence R.
Department of Computer Science, University of California
Santa Barbara, CA 93106, U.S.A.
smithtr@cs.ucsb.edu

Smit, Johanna W.
Universidade de São Paulo/Escola de Comunicações e Artes/Departamento de Biblioteconomia e Documentação
Av. Prof. Lúcio Martins Rodrigues 443, Cidade Universitária, 05508-900, São Paulo, SP, Brazil
Phone: 55-11-3091-4076
Fax: 55-11-3091-4325
cbsdjoke@usp.br

Tálimo, M. de Fátima G. M.
Universidade de São Paulo/Escola de
Comunicações e Artes/Departamento de
Biblioteconomia e Documentação
Av. Prof. Lúcio Martins Rodrigues 443,
Cidade Universitária 05508-900, São
Paulo, SP, Brazil
Phone: 55-11-3091-4076
Fax: 55-11-3091-4325
mfgmtala@usp.br

Tennis, Joseph T.
Information School
University of Washington
Box 352840, Seattle, WA 98195-2840,
USA
jtennis@u.washington.edu

Travieso Rodríguez, Crispulo
Dep. Biblioteconomía y Documentación.
Univ. de Salamanca
Facultad de Traducción y Documentación,
C/P Francisco de Vitoria, 6-16, 37008
Salamanca

Tudhope, Douglas
Hypermedia Research Unit, University of
Glamorgan
Pontypridd, CF37 1DL, Wales, UK
Phone: +1443-482271
Fax: +1443-482715
dstudhope@ glam.ac.uk

Ureña López, L.A.
Departamento. de Informática, Universidad
de Jaén. Avda. Madrid, 35, 23071, Jaén
(Spain)
Phone: +34 953002445
laurena@ujaen.es

Vila Miranda, M. Amparo
Departamento de Ciencias de la
Computación e Inteligencia Artificial.
Universidad de Granada.
C/Daniel Saucedo s/n, 18071, Granada,
Spain
Phone: +34 958244019
Fax: +34 958243317
vila@decsai.ugr.es

Walt, Marthinus S. Van der

Department of Information Science,
Stellenbosch University
Private Bag X1, 7602 Matieland,
Stellenbosch, South Africa
Phone: +27 21 808-2029
Fax: +27 21 808-2117
msvdw@sun.ac.za

Williams, M. Howard
Department of Computing and Electrical
Engineering, Herriot-Watt University,
Edinburgh

Williamson, Nancy
Faculty of Information Studies. University
of Toronto, 140 St., M5S3G6, Toronto,
Ontario,
Canada
Phone: 416 978 7079
Fax: 416 978 5762
William@fis.utoronto.ca

Zapico Alonso, Felipe
Facultad de Biblioteconomía y
Documentación. Universidad de
Extremadura. Alcazaba de Badajoz s/n,
06071 Badajoz, Spain.
fzapalo@alcazaba.unex.es

Zarco, Carmen
PULEVA Salud S.A. Camino de Purchil,
66. 18004 – Granada. Spain

Zins, Chaim
University of Haifa, Israel
chaimz@research.haifa.ac.il

Zozl, Patrizia
Consiglio Nazionale delle Ricerche - ISR,
Roma, Italy

Zumer, Maja
Department of library and information
science and book studies, University of
Ljubljana
Turjaska 1
1000 Ljubljana
Slovenia
Maja.Zumer@NUK.Uni-Lj.Si
Seventh Conference of the International Society for Knowledge Organization (ISKO)
Held in the Palacio de Exposiciones y Congresos of Granada. Granada, Spain July 10-13, 2002

Organizers
International Society for Knowledge Organization (ISKO)
Facultad de Biblioteconomía y Documentación. Universidad de Granada, Granada, Spain

Co-sponsored by:
Ministry of Science and Technology, Spain
ISKO-Spanish Chapter
University of Granada, Spain
Town Council, Granada, Spain

Conference Chair
María J. López-Huertas
Facultad de Biblioteconomía y Documentación. Universidad de Granada (Spain)

In cooperation with
Clare Beghtol, President of ISKO, University of Toronto, Toronto (Canada)

Programme Chairs
María J. López-Huertas, ISKO-Spain & Facultad de Biblioteconomía y Documentación, Universidad de Granada, Spain
Juan Carlos Fernández-Molina, ISKO-Spain & Facultad de Biblioteconomía y Documentación, Universidad de Granada, Spain
Enrique Herrera Viedma, Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, Spain
Óscar Cordón, Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, Spain
Programme Committee
Hanne Albrechtsen, Centre for Human Computer Interaction, Ris National Laboratory Roskilde, Denmark
Clare Beghtol, University of Toronto, Toronto, Canada
Oscar Cordón, Universidad de Granada, Spain
Emilia Currás, Former Spanish ISKO President
Ingetraut Dahlberg, Am Hirtenberg 13, D-64732 Bad König, Germany
Juan Carlos Fernández-Molina, Universidad de Granada, Granada, Spain
Raya Fidel, University of Washington, Seatle, WA, USA
José Antonio Frías, Universidad de Salamanca, Salamanca, Spain
Antonio García Gutiérrez, Universidad de Sevilla, Sevilla, Spain
Rebecca Green, University of Maryland, College Park, MD, USA
Enrique Herrera-Viedma, Universidad de Granada, Granada, Spain
Michèle Hudon, Université de Montréal, Montréal, Canada
María J. López-Huertas, Universidad de Granada, Granada, Spain
Ja. Mcllwaine, University College London, London, United Kindom
Jens-Erik Mai, Royal School of Library and Information Science, Copenhagen, Denmark
Félix de Moya, Universidad de Granada, Spain
Widad Mustafa el Hadi, Université Charles De Gaulle Lille 3, Lille, France
A. Neelameghan, ISI/DRTC, Bangalore, India
Hope Olson, University of Alberta, Edmonton, Canada
Steven Pollit, View-Based Systems Limited, Huddersfield, United Kindom
Gerhard Resthius, University of Amsterdam, Amsterdam, The Nederlands
Rosa San Segundo, Universidad Carlos III de Madrid, Getafe, Sapin
Eugeniusz Scibor, Higher Pedagogical College in Olsztyn, Warsaw, Poland
Alexander Siegel, Informationszentrum Sozialwissenschaften, Bonn, Germany
Dagobert Soergel, University of Maryland, College Park, MD; USA
Paul Solomon, University of North Carolina, Chapel Hill, NC, USA
Elaine Svenonius, University of California, Los Angeles, CA, USA
Martin van der Walt, Stellenbosch University, Stellenbosch, South Africa
Nancy J. Williamson, University of Toronto, Toronto, Canada

Session Chairs
Jack Andersen, Royal School of Library and Information Science, Copenhagen, Denmark
Clare Beghtol, University of Toronto, Toronto, Canada
Oscar Cordón, Universidad de Granada, Spain
Emilia Currás, Former Spanish ISKO President
Ingetraut Dahlberg, Am Hirtenberg 13, D-64732 Bad König, Germany
Juan Carlos Fernández Molina, Universidad de Granada, Spain
José Antonio Frías, Universidad de Salamanca, Salamanca, Spain
Antonio García Gutiérrez, Universidad de Sevilla, Sevilla, Spain
Enrique Herrera-Viedma, Universidad de Granada, Spain
Pedro Hipola, Universidad de Granada, Spain
Peter Ohly, IZ Sozialwissenschaften, Bonn, Germany
Hope Olson, University of Alberta, Edmonton, Canada
Gerhard Resthius, University of Amsterdam, Amsterdam, The Nederlands
Seventh Conference of the International Society for Knowledge Organization (ISKO)
Held in the Palacio de Exposiciones y Congresos of Granada, Granada, Spain July 10-13, 2002

Organizers
International Society for Knowledge Organization (ISKO)
Facultad de Biblioteconomía y Documentación. Universidad de Granada, Granada, Spain

Co-sponsored by:
Ministry of Science and Technology, Spain
ISKO-Spanish Chapter
University of Granada, Spain
Town Council, Granada, Spain

Conference Chair
María J. López-Huertas
Facultad de Biblioteconomía y Documentación. Universidad de Granada (Spain)

In cooperation with
Clare Beghtol, President of ISKO, University of Toronto, Toronto (Canada)

Programme Chairs
María J. López-Huertas, ISKO-Spain & Facultad de Biblioteconomía y Documentación, Universidad de Granada, Spain
Juan Carlos Fernández-Molina, ISKO-Spain & Facultad de Biblioteconomía y Documentación, Universidad de Granada, Spain
Enrique Herrera Viedma, Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, Spain
Óscar Cordón, Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, Spain
Programme Committee
Hanne Albrechtsen, Centre for Human Computer Interaction, Ris National Laboratory Roskilde, Denmark
Clare Beghtol, University of Toronto, Toronto, Canada
Oscar Cordón, Universidad de Granada, Spain
Emilia Currás, Former Spanish ISKO President
Ingetraut Dahlberg, Am Hirtenberg 13, D-64732 Bad König, Germany
Juan Carlos Fernández-Molina, Universidad de Granada, Granada, Spain
Raya Fidel, University of Washington, Seattle, WA, USA
José Antonio Frias, Universidad de Salamanca, Salamanca, Spain
Antonio García Gutiérrez, Universidad de Sevilla, Sevilla, Spain
Rebecca Green, University of Maryland, College Park, MD, USA
Enrique Herrera-Viedma, Universidad de Granada, Granada, Spain
Michèle Hudon, Université de Montréal, Montréal, Canada
Maria J. López-Huertas, Universidad de Granada, Granada, Spain
Jens-Erik Mai, Royal School of Library and Information Science, Copenhagen, Denmark
Félix de Moya, Universidad de Granada, Spain
Widad Mustafa el Hadi, Université Charles De Gaulle Lille 3, Lille, France
A. Neelameghan, ISI/DRTC, Bangalore, India
Hope Olson, University of Alberta, Edmonton, Canada
Steven Pollit, View-Based Systems Limited, Huddersfield, United Kingdom
Gerhard Resthuius, University of Amsterdam, Amsterdam, The Nederlands
Rosa San Segundo, Universidad Carlos III de Madrid, Getafe, Spain
Eugeniusz Scibor, Higher Pedagogical College in Olsztyn, Warsaw, Poland
Alexander Siegel, Informationszentrum Sozialwissenschaften, Bonn, Germany
Dagobert Soergel, University of Maryland, College Park, MD, USA
Paul Solomon, University of North Carolina, Chapel Hill, NC, USA
Elaine Svenonius, University of California, Los Angeles, CA, USA
Martin van der Walt, Stellenbosch University, Stellenbosch, South Africa
Nancy J. Williamson, University of Toronto, Toronto, Canada

Session Chairs
Jack Andersen, Royal School of Library and Information Science, Copenhagen, Denmark
Clare Beghtol, University of Toronto, Toronto, Canada
Oscar Cordón, Universidad de Granada, Spain
Emilia Currás, Former Spanish ISKO President
Ingetraut Dahlberg, Am Hirtenberg 13, D-64732 Bad König, Germany
Juan Carlos Fernández Molina, Universidad de Granada, Spain
José Antonio Frias, Universidad de Salamanca, Salamanca, Spain
Antonio García Gutiérrez, Universidad de Sevilla, Sevilla, Spain
Enrique Herrera-Viedma, Universidad de Granada, Spain
Pedro Hipola, Universidad de Granada, Spain
Peter Ohly, IZ Sozialwissenschaften, Bonn, Germany
Hope Olson, University of Alberta, Edmonton, Canada
Gerhard Resthuius, University of Amsterdam, Amsterdam, The Nederlands
Nancy J. Williamson, University of Toronto, Toronto, Canada
Chaim Zins, University of Haifa, Israel

Session Relators
Grant Campbell, University of Western Ontario London, Ontario, Canada
Oscar Cordón, Universidad de Granada, Spain
Jonathan Furner, University of California, Los Angeles, USA
Enrique Herrera-Viedma, Universidad de Granada, Spain
Hemalata Iyer, University at Albany, USA
Ia. McLlwaine, University College London, London, United Kingdom
Jens Erik Mai, Royal School of Library and Information Science, Copenhagen, Denmark
Widad Mustafa el Hadi, Université Charles De Gaulle Lille 3, Lille, France
Jian Quin, Syracuse University, USA
Richard Smiraglia, Long Island University, NY, USA

Panelists
Peter Ohly, IZ Sozialwissenschaften, Bonn, Germany
Heiner Benking, Consultant, Technician and Engineer, Germany
Emilio Delgado López-Cózar, Universidad de Granada, Spain
Juan Carlos Fernández Molina, Universidad de Granada, Spain
Victoria Francu, Central University Library of Bucharest, Romania
Rolf Goetzke / Katja MruckFQS / qualitative-research.net Free University of Berlin
Debra Hiom, SOSIG ILRT, University of Bristol
Henrike Hoffmann, S I D O S, Swiss information and data archive service for the social sciences, Switzerland
Rosa San Segundo, Universidad Carlos III de Madrid, Spain
Kate Sharp, Project Manager, Biz/ed and Virtual Training Suite for Further Education Institute for Learning and Research Technology, University of Bristol
Jesús Tramullas, Universidad de Zaragoza, Spain
Martin van der Walt, Stellenbosch University, Stellenbosch, South Africa

Acknowledgments
This edition has been partially financed by a grant by the Ministry of Science and Technology of Spain