Contents

International Society for Knowledge Organization.
11th General Assembly 2008. Agenda ...................... 196

Articles

Fulvio Mazzocchi, Melissa Tiberi, Barbara De Santis, and Paolo Plini.
Relational Semantics in Thesauri:
Some Remarks at Theoretical and Practical Levels..... 197

Guglielmo Trentin.
Graphic Tools for Knowledge Representation
and Informal Problem-Based Learning
in Professional Online Communities......................... 215

Jody L. DeRidder.
The Immediate Prospects for the Application
of Ontologies in Digital Libraries.......................... 227

Koraljka Golub, Thierry Hamon, and Anders Ardö.
Automated Classification of Textual Documents
Based on a Controlled Vocabulary in Engineering...... 247

Book Reviews

Murtha Baca, Patricia Harping, Elisa Lanzi, Linda McCrea, and Ann Whiteside (eds.).

Patrick Lambe. Organising Knowledge:
Taxonomies, Knowledge and Organisational Effectiveness.

ISKO News ................................................................. 268

Knowledge Organization Literature
34 (2007) No.4......................................................269

Personal Author Index
34 (2007) No.4......................................................282

Personal Author Index
34 (2007) No.4......................................................282
Contents pages


ABSTRACT: A thesaurus is a controlled vocabulary designed to allow for effective information retrieval. It consists of different kinds of semantic relationships, with the aim of guiding users to the choice of the most suitable index and search terms for expressing a certain concept. The relational semantics of a thesaurus deal with methods to connect terms with related meanings and are intended to enhance information recall capabilities. In this paper, focused on hierarchical relations, different aspects of the relational semantics of thesauri, and among them the possibility of developing richer structures, are analyzed. Thesauri are viewed as semantic tools providing, for operational purposes, the representation of the meaning of the terms. The paper stresses how theories of semantics, holding different perspectives about the nature of meaning and how it is represented, affect the design of the relational semantics of thesauri. The need for tools capable of representing the complexity of knowledge and of the semantics of terms as it occurs in the literature of their respective subject fields is advocated. It is underlined how this would contribute to improving the retrieval of information. To achieve this goal, even though in a preliminary manner, we explore the possibility of setting against the framework of thesaurus design the notions of language games and hermeneutic horizon.


ABSTRACT: The use of graphical representations is very common in information technology and engineering. Although these same tools could be applied effectively in other areas, they are not used because they are hardly known or are completely unheard of. This article aims to discuss the results of the experimentation carried out on graphical approaches to knowledge representation during research, analysis and problem-solving in the health care sector. The experimentation was carried out on conceptual mapping and Petri Nets, developed collaboratively online with the aid of the CMapTool and WoPeD graphic applications. Two distinct professional communities have been involved in the research, both pertaining to the Local Health Units in Tuscany. One community is made up of head physicians and health care managers whilst the other is formed by technical staff from the Department of Nutrition and Food Hygiene. It emerged from the experimentation that concept maps are considered more effective in analyzing knowledge domain related to the problem to be faced (description of what it is). On the other hand, Petri Nets are more effective in studying and formalizing its possible solutions (description of what to do). For the same reason, those involved in the experimentation have proposed the complementary rather than alternative use of the two knowledge representation methods as a support for professional problem-solving.


ABSTRACT: The purpose, scope, usage, methodology, cross-mapping and encoding of ontologies is summarized. A snapshot of current research and development includes available tools, ontologies, and query engines, with their applications. Benefits, problems, and costs are discussed, and the feasibility and usefulness of ontologies is weighed with respect to potential and current digital library arenas. The author concludes that ontology application potentially has a huge impact within knowledge management, enterprise integration, e-commerce, and possibly education. Outside of heavily funded domains, feasibility depends on assessment of various evolving factors, including the current tools and systems, level of adoption in the field, time and expertise available, and cost barriers.


ABSTRACT: Automated subject classification has been a challenging research issue for many years now, receiving particular attention in the past decade due to rapid increase
of digital documents. The most frequent approach to automated classification is machine learning. It, however, requires training documents and performs well on new documents only if these are similar enough to the former. We explore a string-matching algorithm based on a controlled vocabulary, which does not require training documents—instead it reuses the intellectual work put into creating the controlled vocabulary. Terms from the Engineering Information thesaurus and classification scheme were matched against title and abstract of engineering papers from the Compendex database. Simple string-matching was enhanced by several methods such as term weighting schemes and cut-offs, exclusion of certain terms, and enrichment of the controlled vocabulary with automatically extracted terms. The best results are 76% recall when the controlled vocabulary is enriched with new terms, and 79% precision when certain terms are excluded. Precision of individual classes is up to 98%. These results are comparable to state-of-the-art machine-learning algorithms.

These contents pages may be reproduced without charge.
Instructions for Authors

Manuscripts should be submitted electronically (in Word, WordPerfect, or RTF format) in English only to the editor-in-chief, taking into account the merit of the manuscript, and should conform as much as possible to professional standards of English usage and grammar. Manuscripts are received with the understanding that they have not been previously published, are not being submitted for publication elsewhere, and that if the manuscript is accepted for publication the author’s name should not appear on the first page of the text. To protect anonymity, the author’s name should not appear on the first page of the text.

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

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

Reference citations within the text should have the following form: (author year). For example, (Jones 1990). Specific page numbers are optional, but preferred when applicable, e.g. (Jones 1990, 100). A citation with two authors would read (Jones & Smith, 1990). Three or more authors would be: (Jones et al., 1990). A citation with two authors would read (Jones 1990). For example, (Jones 1990). Specific page numbers are optional, but preferred when applicable, e.g. (Jones 1990, 100). A citation with two authors would read (Jones & Smith, 1990). Three or more authors would be: (Jones et al., 1990). When the author is mentioned in the text, only the date and optional page number should appear in parenthesis – e.g. According to Jones (1990), …

References should be listed alphabetically by author at the end of the article. Author names should be given as found in the sources (not abbreviated). Journal titles should not be abbreviated. Multiple citations to works by the same author should be listed chronologically and should each include the author’s name. Articles appearing in the same year should have the following format: “Jones 2005a, Jones 2005b, etc.” Issue numbers are given only when a journal volume is not through-paginated.

Examples:


Footnotes are not permitted; all narration should be included in the text of the article. Illustrations should be kept to a necessary minimum and should be submitted electronically when possible. Photographs (including color and half-tone) should be scanned with a minimum resolution of 600 dpi and saved as .tif files (Tagged Image File Format preferred). Tables and figures should be embedded within the document or, alternatively, saved as separate files with clear instructions indicating their placement in the text. Tables should contain a number and title at the top, and all columns and rows should have headings. All illustrations should be cited in the text as Figure 1, Figure 2, etc. or Table 1, Table 2, etc. Illustrations submitted in hard copy only should be marked to indicate their placement in the text.

Upon acceptance of a manuscript for publication, authors must provide a wallet-size photo and a one-paragraph biographical sketch. The photograph should be scanned with a minimum resolution of 600 dpi and saved as a .tif file (Tagged Image File Format). Advertising

Responsible for advertising: Dr. H.-J. Dietrich, ERGON-Verlag, Grombühlstr. 7, 97080 Würzburg (Germany).

© 2007 by ERGON-Verlag Dr. H.-J. Dietrich.

All Rights reserved.

KO is published quarterly by ERGON-Verlag.
The price is €115.00/ann. including airmail delivery.
The more scientific data is generated in the impetuous present times, the more ordering energy needs to be expended to control these data in a retrievable fashion. With the abundance of knowledge now available the questions of new solutions to the ordering problem and thus of improved classification systems, methods and procedures have acquired unforeseen significance. For many years now they have been the focus of interest of information scientists the world over.

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

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

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

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

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

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

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

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

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

Rates: From 2006 on for 4 issues/ann. (including indexes) € 115,00 (forwarding costs included). Membership rates see above.

ERGON-Verlag, Grombühlstr. 7, GER-97080 Würzburg; Phone: +49 (931) 282084; FAX +49 (931) 282872; E-mail: service@ergon-verlag.de; http://www.ergon-verlag.de

The contents of this journal are indexed and abstracted in Referativny Zhurnal Informatika and in the following online databases: Information Science Abstracts, INSPEC, Library and Information Science Abstracts (LISA), Library Literature, PASCAL, Sociological Abstracts, and Web Science & Social Sciences Citation Index.
I am most pleased to invite you to the 11th ISKO General Assembly which will take place in August 2008 in Montréal, Canada, at the 10th ISKO Conference. All ISKO members are encouraged to attend both the Conference and the General Assembly.

The proposed Agenda is as follows:

1. Opening: Election of General Assembly Chair and the Secretary
2. Approval of & Additions to the Agenda
3. Report of the President
4. Report of the Secretary and Treasure
5. Report of the editor of the journal Knowledge Organization
6. New ISKO Chapters
7. Reports of the Representants of ISKO Regional and National Chapters
8. The Eleventh International ISKO Conference
9. Elections of members for the Executive Committee
   a. Election of Secretary/Treasure
   b. Election of two EC members
10. Any other business

I look forward very much to seeing as many ISKO members as possible at the 10th International Conference in Montréal and at this 11th General Assembly.

María J. López-Huertas, ISKO President.
Relational Semantics in Thesauri: Some Remarks at Theoretical and Practical Levels

Fulvio Mazzocchi*, Melissa Tiberi **, Barbara De Santis ***, Paolo Plini ****

* / ** / *** / ****Institute for Atmospheric Pollution of CNR, Via Salaria km 29, 300 Monterotondo staz., 00015 (RM), Italy,
*<mazzocchi@iia.cnr.it>  **<bds_iia@yahoo.it>  ***<plini@iia.cnr.it>

**Central National Library of Florence, Piazza dei Cavalleggeri, 1, I-50122 Florence, Italy, <tiberim77@yahoo.it>

Fulvio Mazzocchi works as a researcher at the Institute for Atmospheric Pollution of the Italian National Research Council in Monterotondo (RM). He has studied biologic sciences and philosophy at ‘La Sapienza’ University in Rome. He has participated in a number of projects concerned with the design and the implementation of thesauri for the environmental domain, such as EARTh and GEMET. Among his current research interests there are epistemological foundations of and semantics in relation to knowledge organization.

Melissa Tiberi has obtained a degree in philosophy at ‘La Sapienza’ University in Rome. At present, she is working as an external consultant for the National Central Library in Florence, where she is taking part in the development of the Thesaurus of the Nuovo Soggettario. In the past, by making research on the different kinds of semantic relationships and by implementing them in the thesaurus, she has collaborated to the development of EARTh, too.

Barbara De Santis obtained a degree in interpreting and translating (languages: English and German) at the Bologna University. At present she is working as an external consultant for the Italian National Research Council, at the development of the EARTh project, concentrating mainly on multilingual aspects within thesauri.

Paolo Plini was born in 1960 in Rome. He graduated in 1984 in Natural Sciences from the University of Rome. Since 1994 he is researcher at the Italian National Research Council. At present he is the scientific responsible of the Environmental Knowledge Organisation Laboratory of the Institute for Atmospheric Pollution. His main activities are focused on the design and management of EARTh (Environmental Applications Reference Thesaurus) and of other thesauri on specific environmental topics.


ABSTRACT: A thesaurus is a controlled vocabulary designed to allow for effective information retrieval. It consists of different kinds of semantic relationships, with the aim of guiding users to the choice of the most suitable index and search terms for expressing a certain concept. The relational semantics of a thesaurus deal with methods to connect terms with related meanings and are intended to enhance information recall capabilities. In this paper, focused on hierarchical relations, different aspects of the relational semantics of thesauri, and among them the possibility of developing richer structures, are analyzed. Thesauri are viewed as semantic tools providing, for operational purposes, the rep-
A thesaurus is a controlled vocabulary designed to allow for successful information retrieval (IR). It includes different types of semantic relationships that guide indexers and searchers to the selection of the most suitable terms for expressing given concepts/queries (Dextre Clarke 2001). The relational semantics of a thesaurus are concerned with methods to connect terms with related meanings and constituted by the set of meaning relationships. The basic relationships which typify a traditional thesaurus are three: hierarchical, associative and of equivalence. Being functional and not semantic tools strictu sensu, in most cases thesauri do not provide a complete and precise definition of the meaning of terms (Schmitz Esser 1991). The relational structure is designed, in fact, mainly to enhance the information recall performance (Svenonius 2000). Nonetheless, thesauri can still be regarded as (operational) semantic tools in the sense that thesaurus relations are semantic relations and that a thesaurus provides the conceptual structure of a subject field (Hjørland 2007).

A number of scholars have stressed the importance of semantic research in relation to information science (IS), and in particular to its subfield of knowledge organization, which is concerned with "the construction, use, and evaluation of semantic tools for IR" (Hjørland 2007, 369). The kind of meaning understanding can have, in fact, a considerable impact on how knowledge organization systems (KOSs), as a thesaurus, and their relational semantics are designed and implemented. The primary relationships employed in a thesaurus, in fact, although at some levels they reflect certain basic cognitive inclinations of the human form of life (as the one towards classification and hierarchization), are not 'given' as such—and thus necessarily and universally valid—but 'constructed' and defined within a certain (cultural and) theoretical tradition. In some cases, they are even based on assumptions rooted in the centuries of the history of philosophy (Hjørland 2007), as occurs with the notion of genus and species whose origin can be traced back to Aristotle and which is based on an idea of meaning that has been predominant in the Western culture.

A more detailed discussion on such a topic is beyond the scope of this paper and would concern a further investigation on the nature of semantic relations as being mostly theoretical constructs because built within the framework of a cultural form of life (Wittgenstein 1953), this latter being, however, expression of a most basic human form of life, which defines our primary cognitive means and other basic characteristics as being members of the same species. A number of models of conceptualization of the world have been crystallized and with them also certain ways to consider meaningful the relationships between words. In the Western culture, some of these relations (genus-species, synonyms, antonyms, etc.) are common to all knowledge fields. Others are more specific to particular domains (in a thesaurus they can be represented as associative relationships sub-kinds). However, the implementation of any relation always depends on the conceptual and linguistic knowledge of the domain they refer to (in a thesaurus it depends on operational concerns, as well).

Thus, in order to acquire a deeper understanding of KOSs as operational semantic tools, it is important to investigate which theories are behind the principles determining how the relations have to be established. At the same time, it is also important to explore if other theoretical approaches exist and if they can provide useful insights for such issues. A chance to deepen this topic is offered by a new trend in the panorama facing thesauri. In recent years thesauri have entered a larger area of application including knowledge and language engineering. As a consequence, in this new framework and for present and future information retrieval and intelligent processing needs, the thesaurus relational structure is likely to require an enlargement and a refinement of its definition. In order to achieve these goals, a more thoughtful exploration of the theoretical bases that guide its development appears to be necessary.

Analyzing different aspects of the relational semantics of thesauri (the focus will be restricted to the hierarchical relationship) is the subject of this paper, structured as follows. Section I presents the basic roles of relational semantics in thesauri as well as the actual trend towards its refinement. After hav-
ing introduced in section 2 the difference between the instance and the generic relationships, in section 3 we investigate a number of issues involved in meaning representation occurring in thesauri through the classificatory and taxonomic aspects of their relational semantics, such as the criteria upon which the construction of the (logical) hierarchical trees are normally based and the distinction between genus-species and perspective hierarchies. In this framework, what insights may be gained from the perspective of hermeneutics and from Wittgenstein’s notion of language game is explored, too, together with their possible practical implications for the retrieval of information. Section 4 analyzes the partitive relationship and the possibility of its refinement, through a differentiation into distinct subkinds. An overview of existing taxonomies of partitive relations is presented, too. Taking the partitive relationship as a case study, a more general discussion concerning the factors on which the choice of the kind of relations, as well as their implementation depend, is also outlined.

1. Relational semantics in thesauri: its role and possible refinement

1.1 The (general) role of the relational semantics

Thesauri are tools designed for the purpose of improving information retrieval. They are based on a natural language that is transformed, however, by means of certain semantic treatments, into an ‘artificial’ and normalized language where terms are basically monosemic and relations among them are made explicit. Two different semantic structures are used in order to achieve this scope: the referential and the relational semantics (Svenonius 2000). Referential semantics consists of methods to limit the meanings or referents of thesaurus terms: homonyms and polysemes are disambiguated in order to improve precision in IR.

It is through the relational semantics of a thesaurus, that is the object of interest of this paper, that terms are connected to each other when related meanings are identified, devising in this way the relational structure that enhances the information recall performance, although it can also contribute to improve precision by suggesting more specific terms that can refine the search and help to eliminate unwanted information. The network of relations of a thesaurus plays a semantic role since by means of it a further representation of the meaning of each thesaurus term and a structured representation of the general understanding of a subject area are provided. As stated by Soergel (1995, 369), in fact, “a good thesaurus provides, through its hierarchy augmented by associative relationships between concepts, a semantic road map for searching and indexers and anybody else interested in an orderly grasp of a subject field”.

1.2 Trend towards a refinement of the relational semantics

Bearing in mind these important functions of the relational structure, it is then necessary to define the degree of complexity on the basis of which the thesaurus is conceived, in order to ensure its effectiveness for information indexing and retrieval. Methods to measure its richness have already been developed. Examples can range from the number of relation types to more sophisticated indicators, e.g. the ratio of the number of semantic relations and the number of terms which are included in a thesaurus (Van Slype 1976). The traditional thesaurus format—which stems from the more than twenty year old recommendations of the Standard for thesaurus development—has been created to cope with information needs in the library and archival fields (Schmitz Esser 1991).

However, many things have changed and are presently changing (this has been partially reflected in the development of new Standards like ANSI/NISO Z.39.19.2005). Technological advance, which has also brought a larger and differentiated community to search for information on a computer basis, has established a different framework, which requires reassessing prior assumptions and reconsidering whether the existing types of relationships still cope with the current needs of information organization. And actually, a rather widespread opinion is that the traditional thesaurus format is no longer the best-suited means of dealing with these needs. It seems that a richer and hierarchically organized set of relations would be more clearly apt to face them and, as stated by Milstead (2001, 65):

There is reason to expect that provision of semantic relationships in controlled vocabularies will become much more extensive in a future standard, though this does not automatically mean that users will need to be aware of all kinds of relationships in order to use a particular vocabulary.
Despite the general trend towards an expansion of the semantic structure, the outcome of some past experiments comparing systems that incorporate different degrees of semantic structure seems somehow to question the equation more structure - more effectiveness. Besides, in order to evaluate the effectiveness of a semantic structure in IR, other factors should be considered, too, such as the comprehensiveness of the language or the manipulation in retrieval of the subject language (Svenonius 2000). This refinement is necessary to enhance thesaurus suitability for uses in the artificial intelligence (AI) and the Semantic Web environments, as well as to increase possibilities for IR. In particular, AI applications are creating a demand for more elaborated KOS able to ensure higher expressive capabilities in order to allow inference (Dextre Clarke 2001). In such a setting, the traditional relational structure is considered insufficiently detailed and lacking of a well-defined semantics. “All the well-know relationships are fuzzy in most thesauri. We could afford to allow them to be fuzzy as long as their only purpose was to achieve the desired degree of order in our documents, which is a modest requirement compared with what we need for Language and Knowledge Engineering” (Schmitz Esser 1991, 145).

Hence, along gaining a higher (conceptual and lexical) user interaction with the KOS in that the refinement of the relational semantics might improve query formulation and subject browsing, examples of new applications for which such refinement is advocated include supporting automated processing; query expansion; RDF representations of thesauri for the Semantic Web; and interoperability among different KOSs (Soergel et al., 2004; Tudhope et al., 2001).

Finally, the adoption of more expressive semantic relations is advised also to improve the degree of internal structural consistency. In many cases, in fact, the standard set of relationships has not been consistently applied (for instance, many links, labelled as hierarchical, could be best resolved through an associative relationship). For some authors, this is exactly a consequence of the fact that thesaurus relationships are not provided with a precise semantics (Soergel et al., 2004).

Some advanced thesauri are developing or have already included—mainly in the medical domain as UMLS or MeSH—richer sets of semantic relationships. A further example is the Italian CNR’s EARTh project (Mazzocchi & Plini, 2005). Other projects, such as the FAO’s AGROVOC, are instead more concerned with the reengineering of thesauri into ontologies. They aim at developing an enriched set of relationships—the latter would be explicitly labelled and applied with specification of rules and constraints—on the basis of a more fully concept-oriented organizational model, where concepts are regarded as independent from and preceding their designation (Soergel et al., 2004). Indeed, the approach towards building thesauri with an extended relational structure partially converges with the idea and work behind ontology development. An investigation on ontologies, however, is not the focus of the present paper, even though a number of assumptions that are normally associated with them are part of the discussion.

The idea of developing thesauri and other KOSs with a more precise and rich semantics, or of using formal logic methods, and employing a notion of concept as if it were an a priori entity, can somehow be viewed as expressions of the same theoretical point of view, based on logical positivism. What is searched for is creating the conditions for an unambiguous interpretation of terms and relationships mainly to make KOSs suitable for AI applications. According to Svenonius (2004, 585):

The knowledge representations resting upon the epistemological foundations of logical positivism in its operationalist and representational approaches to meaning are … formalized to a greater degree and as such are simpler, more uniform, and relatively free from subjective interpretation. The objectivity they provide through definitional rigor is essential for automated applications in retrieval.

This idea of objectivity, however, conflicts with the fact that meanings and semantic structures in KOSs are always established within a given horizon (reflecting certain theoretical views and applied to specific knowledge domains and operational contexts).

While, of course, the choice to reduce the complexity of reality for operational purposes can be made, and attempts of narrowing it down to such an extent that it becomes manageable are not rare in the AI tradition, a better refinement and specification of relations or the adoption of a logicist view of semantics does not eliminate as such the issues posed by this complexity.

The role played by human judgement in such a task and the multiplicity of different contexts in
which all of this can occur cannot, in fact, be ignored. And this is something that we will try to demonstrate throughout the whole paper, with special focus, though, on the hierarchical relationships.

2 An introductory note on the hierarchical relationship in thesauri

The hierarchical relationship connects pairs of terms when the scope of the broader term (BT) fully includes the scope of the narrower term (NT). Generally speaking, the purpose of the hierarchical relationship is to provide a semantic tree pathway, which can be useful both as a tool for semantic control and specification—the meaning of each term is, in fact, (partially) identified by its position within the tree—and as a navigational aid, by offering users the possibility to choose the terms to employ, when referring to a certain concept, among a range situated at different levels of specificity (Dextre Clarke 2001). This relation comprises the following three different kinds: generic, instantial and partitive. In a restricted number of thesauri they are distinguished as follows:

<table>
<thead>
<tr>
<th>BTG/NTG:</th>
<th>generic</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTP/NTP:</td>
<td>partitive</td>
</tr>
<tr>
<td>BTI/NTI:</td>
<td>instantial</td>
</tr>
</tbody>
</table>

The next section will first introduce the generic and instance relationships. Then, a discussion about the main features of the generic relation and a comparison with perspective hierarchies will follow. Special emphasis will be placed on how any given classification or hierarchization of a term depends on which of its conceptual features are made salient in the light of a given perspective. Section 4, instead, will analyze the partitive relationship.

3 The generic and instance relationships

The generic relationships—named also inclusion, subsumption or hyponymy—connects a genus with its species (e.g., animals—mammals). An important property of this relation, also used as a criterion for its identification, is the inheritance of properties: any attributes of the genus (hypernym) must also be attributable to the species (hyponym). In this sense, the meaning of the hyponym derives from the meaning of the hypernym, plus some additional features. Chaffin et al. (1988) distinguished four kinds of inclusion according to the type of concept involved: natural object-kind; artefact-kind; state-kind; and activity-kind. In the instance relationship the narrower terms are nor parts neither types, but individual instances of the broader terms. In a thesaurus, this characteristic of individuality is expressed through a proper name (e.g., deserts—Sahara desert).

At this stage, the distinction between generic relationship and instantiation seems clearly stated. Nonetheless, Milstead (2001) has emphasized that in the standards for thesauri there is no method used to determine the genus-species relationship that could not be applied also to the instance relationship. For example, the ‘all-and-some’ test—which is used to assess the validity of the generic links (ISO 1986)—can be applied to both cases (if grammatical differences in number are admitted). The same is true also for ‘is a’ attribution:

1a. All mammals are animals
1b. Some animals are mammals
2a. A mammal is an animal
2b. The Sahara desert is a desert

All of this may also lead to conceive the instance relationship as a variant of the genus-species relationship. However, unlike the generic one (concept-to-concept relationship) the instance relationship points to a change of ‘logical level’ (individual-to-concept relation).

3.1 Associative, perspective and logically-based hierarchies

The hierarchical relationship, and particularly the generic kind, is perhaps the most important within a thesaurus and its proper application plays a key role in ensuring the quality of a structured vocabulary. But can we estimate such aptness in an abstract sense? It is true that in many thesauri this relationship has been implemented in quite an inconsistent way, often resulting in unpredictable semantic structures (Dextre Clarke 2001).

As mentioned before, a higher degree of rigour is thus advocated to improve the level of structural consistency. Nonetheless, different contexts may require different solutions, each having its own implications. Furthermore, it is of the utmost importance to investigate the underlying assumptions that the generic relationship, on which basis hierarchical trees
are built, entails not only to deepen our understanding of it, but also to have the chance to critically analyze these assumptions in the light of a comparison with alternative models.

3.1.1 RT-kind version of hierarchy

Many existing thesauri have labelled as hierarchical relations between terms not belonging to the same conceptual category. An example of it can be found in the GEMET thesaurus where the term Recycling ratio (a parameter) is considered to be a Narrower Term of Recycling (an operation). Relationships like this have been established according to a definition of hierarchy that is of a ‘pragmatic’ nature and oriented towards the function of the search process: “Concept A is broader than concept B whenever the following holds: in any inclusive search for A all items dealing with B should be found. Conversely B is narrower than A” (Soergel 1974, 79).

Using such a version of the hierarchical relation can be useful to manage certain databases. But if it may somehow function efficiently at local levels, i.e. in a specific operative context, in a different and wider framework, this choice may result unsatisfactory, since a so-developed hierarchy would suffer lack of consistency with other structures, not being conform to the standard thesaurus format. Moreover, confusion may also arise if RT-kind (associative) hierarchies, like the above example, are labelled in the same way as the genus-species relation (or in any case as a hierarchical kind).

3.1.2 Genus-species and perspective hierarchies

In developing the thesaural relational structure, and thus hierarchies, Foskett (1980) emphasized the importance of the logical perspective: a thesaurus would benefit if the choice of terms and relationships reflected the logical structure of a subject field, instead of being a scarcely systematized gathering of terms extracted from the literature. Other authors as Maniez (1988) stressed that the usefulness of logical relationships should be subordinated to the purposes of information indexing and retrieval. Svenonius (2000), for her part, underlines the distinction between genus-species and perspective hierarchies. In a more general sense, this distinction, taken up by a number of thesaurus standards, is expressed as being between paradigmatic/a priori relations—e.g., genus-species and syntagmatic/a posteriori ones—among them, perspective hierarchies. The genus-species relationship is viewed as logically-based, definitionally true and functioning context independently. Besides, corresponding to the logical relationship of inclusion, it has been defined in terms of the properties of reflexivity, antisymmetry and transitivity.

Conversely, perspective hierarchies are regarded as functioning more contingently in given empirical contexts and depending on the point of view. Normally, they are not provided with the same logical properties of the generic hierarchies. They express, in fact (Svenonius 2000, 164):

Points of view or aspects from which an object or concept is regarded. In many discipline-based classifications, the point of view is the knowledge domain in which the object or concept is located …. The genus-species relationship limits a rat to being a rodent; a perspective relationship allows it to be an agricultural pest, an experimental animal, and so on.

Thesaurus standards argue that relationships to be included in a thesaurus should be a priori rather than a posteriori. However, the genus-species and the perspective relationships can have different functions and, in defining which hierarchical relationships a thesaurus has to be made of, different factors should be taken into consideration, including the characteristics of the vocabulary to be structured and the purpose for which the relations are intended in retrieval.

Concerning the first point, Svenonius (2000 and 2004), for example, in terms of hierarchy, considers a stricter logical ordering as particularly apt to structure terms whose meanings are somehow more fixed, e.g. scientific terms, whereas she regards perspective hierarchies as more suitable to represent polysemantic and vague lexicons, as is mostly the case in social sciences. Regarding the second aspect, the genus-species relation, being logically based, is valuable, for example, for search broadening and narrowing as well as for retrieval strategies playing on inheritance properties. Perspective hierarchies, instead, are not suitable for these applications. Their added value in IR consists of providing contexts that elucidate from which point of view is a term being considered. In this way, they can assist in navigation and are apt for the disambiguation of multireferential terms (Svenonius 2000).

Perspective hierarchies are used by classifications such as the Dewey Decimal Classification (DDC). The term ‘Insect’, for example, while it can be located only in a single genus-species hierarchy (BT:...
'Arthropoda'), it can instead pertain to several perspective hierarchies according to the points of view from which its meaning is regarded: an insect can be viewed, for example, as an agricultural pest, a disease carrier, etc. (Svenonius 2000 and 2004). In the EARTH thesaurus, the idea of multiple thematic classifications of terms as a complement of placing them into the genus-species tree has been developed on a similar basis (Mazzocchi & Plini, 2005).

It should be noted that terms linked by perspective hierarchies belong to the same conceptual category. Yet, being these links based on a situated perspective, they are not amenable to the 'all-and-some' test and thus, according to a strict application of the standards, not accepted as a valid hierarchy. To explain this, ISO 2788 mentioned as an example 'Parrots BT Birds', which is invariably a true (generic) hierarchy, and thus compatible with the all-and-some test, and 'Parrots BT Pets', that, however, is not (being a perspective hierarchy), since some Pets are Parrots, and only some Parrots are Pets. Yet, if this is mostly true, there may be special cases or particular circumstances where this does not apply. For example, in the restricted context of a specialized thesaurus on domestic animals, Parrots as NT of Pets can be, instead, accepted.

Anyway, despite special cases, being perspective hierarchies somehow context-dependent, it seems that only genus-species hierarchies have the potential to provide the basis for a more consistent application throughout different systems.

3.1.3 The all-and-some test

Indeed, this matter is more complex than it appears. A couple of criteria are normally used to determine genus-species hierarchies. First, terms have to belong to the same conceptual category. This is a necessary (but not sufficient) condition to ensure that a hierarchy is logically based. Both the logical and perspective hierarchies are compatible with it, but (normally) not the RT-kind hierarchy.

The other criterion is compatibility with the all-and-some test. In this latter, Fisher (1998, 20) has recognized the extensional definition of subsumption:

Informally, it is said there that concepts are taken as classes which have members, and that for a genuine narrower concept [all] its members must also be members of the broader concept while for the broader concept only [some]

of its members must also be members of the narrower concept.

It should be said, however, that if on the one hand its usefulness is undeniable, on the other this test seems to present a number of issues that still need to be addressed. For example, the test does not discriminate which levels of a genus-species tree are linked when establishing a hierarchy. 'Parrots BT Birds', 'Parrots BT Animals' and 'Parrots BT Organisms' are all validated as hierarchies, since all parrots are birds, animals and organisms. But, of course, they encompass a different degree of (conceptual) information.

3.1.4 The intentional definition of the generic relationship and its historical predecessor

Naturally, the genus-species relationship may also be described on the basis of a representation of terms/concepts as sets of attribute values or features. We proceed from superordinates to subordinates, which contain all the attribute values of the former, by means of the addition of further key conceptual features (Fugmann 1993). In this formulation, Fisher (1998) has recognized a form of the intentional definition of subsumption. Of course, as concepts become more specific they will also correspond to smaller classes of referents.

In order to better clarify this scheme, it might be helpful to briefly refer to the philosophical tradition from which it derives. Broadly speaking, the origin of the notions of genus and species in the history of the Western thought can be traced back to Plato's and Aristotle's philosophies, whereas the representation of a series of subsequent genus-species links, that starting from a top level (categories) go down to the ultimate or infima species—which in turn are superordinate to the individuum—through a vertical taxonomic structure, was firstly conceived with the Porphyrian tree.

The crucial notion for the establishment of the genus-species relationship is that of specific differentia, which represents the key distinctive element differentiating a species from all others sharing the same genus (co-hyponyms). For example, the category 'substance' with the specific differentia 'material' becomes the subordinate genera 'body', while with the differentia 'immaterial' it becomes 'spirit.' The tree in figure 1 derives from adding, along different hierarchical levels, differentiae to the first of the ten Aristotle’s categories, substance. Even though Aris-
totle never puts it in this way, by means of the same method analogous trees are expected to be developed from any of the other categories (quality, quantity, relation, where or place, when or time, position, having or state, action or operation, passion or process). According to some authors (Girgenti 2004, introduction to Porphyry’s Isagoge), the genus-species tree can be navigated both in an upward direction—ascension, according to a logical point of view—or in a downward direction—declination, based on an ontological perspective.

The same notion of differentia plays a key role also in defining. A classic example is the definition of man (human) as a ‘rational animal.’ The parts of this definiens are ‘animal’, the proximate genus that incorporates within its range of meaning all the essential elements of the superordinate genera and ‘rational’, the specific differentia distinguishing man from all other animals. Listing all the differentiae, ‘human’ is defined as ‘rational sensitive animate material substance.’

Summing up, in a hierarchical arrangement obtained in this way, two items are most relevant: the mechanism of conceptual feature addition (the lower level is always a subclass of the higher one) and the key differentiating character of the added conceptual features. For Aristotle, such a method reflects, on the logical and language planes, a principle that operates on an ontological level with the purpose of identifying the distinctive features of things. Should the latter be adopted, the problem is then how to put it into practice, also considering that our highly structured contemporary knowledge systems seem to be developing more on a horizontal and sectorial plain, than on a vertical level, as a univocal unfolding from an Ur-structure.
More generally, the possibility itself of accessing on a rational level the 'meta' point of view—i.e., the fundamental 'place of observation' where the ontological order is unveiled—has become, from an epistemological point of view, questionable and thus, together with it, also the chance to separate, in a final and objective way, what is essential from what is accidental and to develop that 'unique' genus-species tree, which derives from the further addition of specific differentiae to the top categories.

According to Eco (1983), also Aristotle in some of his works, such as De partibus animalium, recognizes at another level the possibility to develop multiple trees, that could be complementary among themselves, according to different perspectives. Given the impossibility to univocally distinguish accidental from distinctive features, such characteristic of distinctiveness can, in Eco’s view, be acquired only in relation to a situated perspective (e.g., the classificatory or definitory problem in question).

Contemporary biological systematics and taxonomy provide an interesting example of synchronic copresence of different theoretical approaches. The classic Linnean approach—arranging organisms by their morphological similarities—and cladistics (or phylogenetic systematics)—where living beings are classified on the basis of their order in branching in an evolutionary tree—coexist and may also be used in a combined way to obtain further information. Different (theoretical) perspectives can, thus, lead to focusing on a diverse set of characteristics. But they need not necessarily be regarded as being in opposition. There may be cases in which they provide complementary information, useful in obtaining a more complete picture of the matter.

3.1.5 Classification as interpretation

Broadening the perspective, this latter position may (partially) be related to the notion of interpretative horizon as developed, in Gadamer’s work, in the framework of contemporary hermeneutics. Such a notion, in fact, has mainly been used to explain the historicity of human understanding, yet in a more general way it can be regarded as the range of vision including “everything that can be seen from a particular vantage point” (Gadamer 1976, 302). In opposition to an objectivistic and universalistic view, the idea of 'classification as interpretation' acknowledges the fact that any classificatory act is always made from a delimited horizon, which determines how classification is conceived and undertaken and, thus, within the limits of certain basic constraints, which aspects of an item (term or object) are made salient.

In information science, Hjørland and Nissen Pedersen (2005) have developed a theory of classification for IR (that by extension can be applied to hierarchization) somehow reflecting this principle and that has been summarized by Hjørland himself (2007, 373) as follows:

Classification is the ordering of objects (or processes or ideas) into classes on the basis of some properties. (The same is the case when terms are defined: It is determined what objects fall under the terms) .... The properties of objects [which are portrayed in the conceptual features of the terms used to name such objects] are not just 'given' but are available to us only on the basis of some descriptions and pre-understandings of those objects [although these still have 'objective' properties] .... Description (or every kind of representation) of objects is both a reflection of the thing described and of the subject creating the description .... The selection of the properties of the objects to be classified must reflect the purpose of the classification. There is no ‘neutral’ or ‘objective’ way to select properties for classification because any choice facilitates some kinds of use while limiting others .... Any given classification or definition will always be a reflection of a certain view or approach to the objects being classified.

Regarding classification as interpretation means to acknowledge the fact that we always act from a classificatory horizon (Paling 2004). This notion, however, needs to be further explained and this can be done by indicating its possible constitutive elements. First, it comprehends the ontological and epistemological meta-assumptions that provide the 'lens' through which we look at the world (Kuhn 1970) and the way in which they are reflected in the scientific activity. For example, positivism and instrumentalism or hermeneutics have different views of the (same) world and, accordingly, lead to different conceptions of classification and hierarchization, too. Secondly, it includes the domain to which the classification is referring. As stressed in their theory by Hjørland and Nissen Pedersen (2005), criteria for classification are (usually) domain-specific, since different domains may need different descriptions and
classification of items in order to meet their specific purposes.

For example, ‘benzene’ can be described and defined in several different ways depending on the discipline or context in which it is considered. Chemists, of course, emphasize its structural properties in being precursor of a class of chemical compounds. Yet, physicists may focus on other properties and see it as a volatile and inflammable. Other descriptions can emphasize its possible effects—biologists may consider its toxicity and the different routes through which it can enter an organism—or employments—engineers would consider it as a fuel for combustion engines (Fugmann, 1993). Furthermore, the fact that within the same domain conflicting paradigms and views can coexist should also be taken into consideration (Hjørland 2007, 385): “in every domain, there exist different theories, approaches, interests, or ‘paradigms’, which also tend to describe and classify objects according to their respective views and goals.”

Finally, the purpose of classification plays a role in determining the classificatory horizon, too. In fact, even if a domain can be viewed in terms of a common paradigm, different practical concerns may lead to different choices in establishing classificatory and hierarchical structures.

3.1.6 Possible insights from the language games theory

In this context, we believe that the notion of language games (Sprachspiele) can play a significant role and be relevant for IS issues, too. This notion has been introduced by Wittgenstein (1953) to explain the multiplicity of language practices that occur within a language. Language does not consist, in fact, of a single unified game. It is regarded, instead, as a collection of multiple and indefinite games. The basic assumption of this theory is that the meaning of a word should not be regarded in terms of its referent, but of its use. Speaking language is a social action. To know the meaning of a word means to know how to use it as part of an activity, within the framework of a particular language game and its rules.

Wittgenstein has introduced also the notion of family resemblances. Considering several possible and different Sprachspiele, the instances of the use of a word do not (necessarily) share a common denominator or essence (as it is, instead, assumed in class inclusion). They are ‘peripherically’ linked through family resemblances, being similar but each in a different manner, like members of a family (where some may have the same eyes, others the same form of mouth or chin, but without a single feature that necessarily all share).

Following this theoretical approach, it is clear that, having language and meaning the above characteristics, they should not be confined to the rules of a particular language game. Should a deeper investigation still be required, this has a number of important implications with respect to the idea of hierarchical arrangement (in general and applied to a thesaurus) and to a number of other issues. As stated by Svenonius (2004, 578):

Subscribing to the concept of language games entails subscribing as well to the position that knowledge representations are not descriptive of things and relations in the real world; rather they are descriptive of linguistic behavior. The use of knowledge representations to organize information is one kind of language game, one kind of linguistic behaviour.

Besides, linking again the main point to what has been said in the previous paragraph, it could be affirmed that each field of knowledge, which has its own set of conceptualisations, has also its particular language games with specific rules (although this does not mean that they cannot share common elements). Meaning of words can, therefore, change (at least partially) from one domain to the next: “the meanings of words—and, thus, words used to name subjects—are in part fixed and, in part, variable. The variable part assumes its value by being contextualized within a system of concepts” Svenonius (2004, 581).

Further considerations would be needed to investigate whether a hierarchy of conceptual features is possible, if some of these features cannot be ‘cancelled’ (without causing the total alteration of the associated meaning) and what their nature is. The meaning of a term has, in fact, also a more stable part, that is likely to be maintained also after a major paradigm shift or along different domain-based viewpoints. Coming back to the example of ‘benzene’, all the listed descriptions share a common premise: benzene, first of all, is a ‘substance’ (that can have toxic effects, be used as fuel, etc.). Similarly, although diverse taxonomizations of a certain kind of animal may be possible (see note 5), none of them questions its recognition and classification at a higher level as an animal. These features, thus, provide a more stable background while modifications occur mostly at a foreground level.
Furthermore, in a given historical period, being expression of the dominant view, certain semantic relations (and then those conceptual features on which their establishment is based) appear to be more ‘stable’ and can be (extensionally) validated by the all and-some test. For example, according to the taxonomy of the scientific discipline which is interested in studying it (chemistry), benzene \(<_{is\_a}>\,\text{‘organic aromatic substance’}\) and this ‘always’ holds. But this is not always the most important aspect in terms of application. In a nature conservation thesaurus, it might be more useful to represent the meaning of benzene as a ‘pollutant’ rather than as an ‘organic aromatic substance’. It is, however, true that this kind of relationships, in virtue of the stronger consensus sustaining their institution, can (at least) provide a basis to ensure a certain degree of compatibility and interoperability among different systems.

Of course, not all the words convey meaning in the same manner. Some of them have more variable meanings, i.e. more dependent on the context, than others. For example, words used in the social sciences are regarded to have more variable meanings, whereas words used in science as having more fixed meanings. But this is only partially true. Not only, in fact, meaning of scientific words changes along history in correspondence of paradigm shifts (Kuhn 1970). The idea that, in a given historical moment, science is a knowledge system based on universal conceptual structures and that words used in scientific discourses have one and the same meaning in all disciplinary domains has been questioned by part of the XXI century epistemology. Kuhn (2000), for example, regards each discipline or community of practitioners of a certain scientific field as bearing its own set of conceptualizations, crystallized in a particular lexical taxonomy, in the frame of which terms acquire specific meanings. This implies that for a (restricted) number of terms meaning changes along different disciplinary fields (local incommensurability).

Evidently, this fact can be particularly relevant for the design of the hierarchical arrangement of scientific thesauri whose subject field is multidisciplinary (as those devoted to ‘environment’). Moreover, the fact that in a given field of knowledge, different theoretical views can exist simultaneously, providing different descriptions of objects and interpretations of the meaning of terms, although less evident (and also less agreed upon) may be applied to scientific disciplinary areas, too (see also note 5).

Thus, in all cases, concepts are not a priori (and as such universal) entities, but should be regarded in the context of a given conceptualization system in which they are embedded. The meaning of words, including those that are part of scientific vocabularies, should be understood according to the rules of the language games they belong to. The same word can have (slightly or significantly) different meanings according to its use in diverse language games, which can pertain to different knowledge fields or to different theoretical views inside the same domain.

3.1.7 Implications for the retrieval of information

Both principles based on a hermeneutic perspective and the language games theory have practical implications for the retrieval of information (based on the use of a thesaurus). Many databases contain, in fact, documents that have been produced in different subject fields and, when within the same domain, sometimes according to different theoretical perspectives. Meaning, however, cannot be defined by examining the documents of a literature as such. Documents should rather be seen as a means to access the conceptual structure of a given knowledge field and the language games that it encloses.

Words (used in documents), in fact, pertain to given language games. Each paradigm within a given domain (of which it embodies the ‘cognitive’ authority), specifies the basic rules of the use of any term and, then, its meaning. If searchers, as is actually the case, look for concepts (contained in documents) as defined in subject fields and their literatures, semantic tools such as thesauri should be able to represent—by means of their hierarchical arrangement and other relations—the meaning of words consistently with how these are defined in the language games of such domains. The retrieval of information would, in fact, be facilitated if a subject field represented in the documents of a database had such documents indexed and searched by means of words used in accordance with the (domain-based) language games they refer to (Andersen & Christensen 1999).

In particular, users should be made aware of the possible different views on the meaning of words (as occurs in different language games) and, thus, of all the possible different views on a given topic (that can focus on as many aspects of it) which may be useful for them (Hjørland 1998). As underlined by Hjørland (2007, 389), while attempts at standardizing terminology can cause the removal of some of these views, “a precondition for designing quality
KOS is that the designer knows the different views and is able to provide a reasonably informed and negotiated solution."

Of course, a thesaurus has its own language game, too, whose rules are basically oriented towards the achievement of a semantic univocity for operational purposes. However, there are a number of devices that can be used in a thesaurus to represent the different aspects of the semantics of terms and (wherever necessary) to disambiguate them. One of these is the coupled use of genus-species and perspective hierarchies, in order to exploit the different functions that they could have. As already mentioned in 3.1.2, perspective hierarchies can provide additional views about the semantics of a term (or the aspects of a given topic) and can be used for disambiguation purposes, while ‘all-and-some’ hierarchies can also provide a shared basis to make different KOSs more compatible and interoperable.

4. The partitive relationship

This section deals with the partitive relationship. A number of taxonomies organizing it into subclasses are also presented, followed by some remarks on the role played by ‘interpretation’ in implementing these relations (and semantic relations in general) to satisfy the needs of different conceptual contexts and empirical circumstancies.

In the partitive relationship (also named meronymy) the narrower terms are parts of the broader ones. In linguistics, a number of test-frames are used to detect it, such as ‘an X is a part of a Y’ (or inversely ‘a Y has an X / Xs’), but none of them seems to provide an unambiguous indicator of it, since they can also be used to express non-meronymic relationships (Cruse 1986).

Furthermore, which basic properties (among reflexivity, antisymmetry and transitivity) may be ascribed to this relationship is still a debated topic (Iris et al. 1987; Winston et al. 1986). As a rule, thesauri standards regard only four types of this relation as hierarchical: those taking place among parts of the body; organizational structures; geographical locations and disciplines or fields of knowledge. All other cases are classified, instead, as associative relationships, even though exceptions may be accepted in specific subject areas (ISO 1986). The partitive relationship is, thus, not restricted to material objects and should be viewed as a collection of different subkinds (Iris et al., 1988). Yet, no consensus has been reached on the identification of such subkinds, nor has on the linguistic patterns that express them.

4.1 An overview of existing taxonomies of partitive relations

A number of interesting studies have been undertaken in different knowledge fields, such as linguistics, logic and cognitive psychology, in order to develop a taxonomy of partitive relationships. Mostly, they focus on the degree of differentiation of the parts and on their role with respect to the whole. Despite their different origins and aims, the outcome of these studies provides useful insights also for a refinement of this relationship in thesauri.

Perhaps the most influential taxonomization is by Winston et al. (1987), based on experimental data and on a psychological perspective. Winston and his co-workers distinguish six subtypes on the basis of the values of three relational elements, which summarize the attributes of the relationships:

1. Functionality (functional/non functional): parts are/are not in a specific spatial or temporal position with respect to each other, which sustains their functional role with respect to the whole.
2. Degree of similarity (homeomerous/non homeomerous): parts are similar/dissimilar to each other and to the whole to which they belong.
3. Spatial cohesion (separable/inseparable): parts can/cannot be physically separated from the whole.

<table>
<thead>
<tr>
<th>Subtypes</th>
<th>Examples</th>
<th>Functional</th>
<th>Homeomerous</th>
<th>Separable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integral object-component</td>
<td>Cup-handle, Linguistics-phonology</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Collection-member</td>
<td>Forest-tree</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Mass-portion</td>
<td>Salt-grain</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Object-stuff</td>
<td>Bike-steel</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Activity-feature</td>
<td>Shopping-paying</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Area-place</td>
<td>Desert-oasis</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
</tbody>
</table>

*Table 1. Winston et al.’s taxonomy of the partitive relation*
This scheme has already been integrated in some advanced thesauri, e.g. in the project for the development of an environmental thesaurus—EARTH (Environmental Applications Reference Thesaurus).

Together with the description of each relation of the Winston et al.’s scheme, in order to have a look at some results of this implementation, we have listed a number of demonstrative partitive cases extrapolated from EARTH’s environmental (and closely related) terminology.

**Integral object-component**

It takes place between a whole (an ‘integral object’)—which presents some kind of patterned organization or structure—and its components. These latter are also patterned and generally bear specific structural and functional relationships to one another and to the whole of which they are parts. Integral objects consist both of things having an extensive dimension, such as physical things (e.g., natural objects or artefacts), and things whose parts are not extensively contained in their wholes, such as abstract objects and organizations. Due to this reason, a further differentiation in subtypes might still be planned. Accordingly, in the EARTH thesaurus this relation is expressed as follows: `<has_component/is_component_of>`, used for material objects—these include, for example, biological systems (cells, anatomical structures, plants) and, among artefacts, instruments, installations and buildings—and their parts; and a second expression, which however still needs to be defined (for the time being the generic `<has_part/is_part_of>`) to be used, instead, to express the relation between abstract entities, as for example disciplines, and their ‘parts’.

- **Cell <has_component> Cell membrane**
- **Cardiovascular system <has_component> Heart**
- **Electric vehicle <has_component> Electric engine**
- **Ecology <has_part> Land ecology**

**Collection-member**

It records membership in a collection. This relationship does not require that members have a given structural organization or carry out a particular function in relation to each other and to the whole. Collection-member has some similarity to (and can consequently be confused with) the relationship of inclusion since both involve membership of individuals in larger sets. Nevertheless membership in a class (genus) is determined by similarity to the other members (species) based on a set of intrinsic properties. Membership in a collection is instead defined on the basis of characteristics that are extrinsic to the individual members, such as spatial or temporal proximity or a social connection. Chaffin and Herrmann (1988) distinguish three subkinds of this relationship: `<group-member>`, `<member-collection-member>` (e.g., tree-forest, fleet-ship); and `<organization-unit>` (e.g., army-battalion). Up to now, in EARTH `<collection-member>` has been applied to connect material objects and is expressed by `<has_member/is_member_of>`.

- **Flora <has_member> Plants**
- **Game <has_member> Game species**
- **Car population <has_member> Car**

**Mass-portion**

Portions are homeomerous parts of physical objects or masses since every portion is similar to the others and to the whole. They have arbitrary boundaries and lack functional relation to the whole. They should also be distinguished from ‘pieces’ that originate, for example, from the destruction of an object and, unlike portions, are not always homeomerous. In Cruse’s words (1986, 158) “The contrast between parts and pieces is potentially operative even with highly integrated wholes such as animal bodies: there is a clear difference between such a body hacked to pieces, and one carefully dissected into its parts”. Chaffin and Herrmann (1988) make also a distinction between `<mass-measured_portion>` (e.g., pie-slice) and `<mass-natural tiny piece>` (e.g., salt-grain). Furthermore, they include also `<measure-unit>` (e.g., mile-yard) as a third subkind. In EARTH, so far it has had a quite limited application and is expressed by `<yields_portion/is_portion_of>`.

- **Land <yields_portion> Parcel of land**

**Object-stuff**

This relation links an object to the substance or material from which the object is naturally made or manufactured/created. It differs from the object-component relationship in that the stuff of which an object is made cannot be physically separated from it without altering its identity. Chaffin and Herrmann (1988) distinguish `<mass-stuff>` (e.g., trash-paper) from `<object-stuff>` (e.g., lens-glass). These authors, like others such as Ahmad and Fulford (1992) and Iris et al. (1988), do not regard this relationship as partitive. It can, in fact, be considered also as a kind of associative relationship, as has occurred, for example, in EARTH where it is expressed by `<consists_of/is_matter_of>`.
Road <consists_of> Asphalt
Can <consists_of> Tin
Bicycle <consists_of> Aluminium

Activity-feature
It points to the relation focused on those parts—
phases, stages, discrete periods, features, etc.—that
form, in a structured manner, a process or an activity,
which constitutes the whole. Chaffin and Herrmann
(1988), who do not include it among partitive kinds,
distinguish process-phase (e.g., growing up-adoles-
cence), continuous activity-phase (e.g., cycling-pedal-
ing), and discrete activity-phase (shopping-buying).
In EARTh, this relationship has been applied to
(mostly natural) processes and to (social and other
above mentioned Chaffin & Herrmann (1988), distin-
guish a set of subkinds by using relational elements
that do not coincide with those of Winston et al.. Iris
et al. (1988), propose a classification founded on four
basic models. Three of them (the functional compo-
nent; the segmented whole; collection and members) are
similar to the first three Winston et al.’s categories,
whereas the other (sets and subsets) resembles the no-
tion of class-inclusion. Another comparable list has
been proposed by Gerstl & Pribbenow (1995), who
identify kinds induced by (mass/quantities, collection/
elements and complex/components) or independent of
(segments and portions) the compositional structure.
Finally, Cruse (1986) classifies the partitive rela-
tionship according to quantificational differences.

<table>
<thead>
<tr>
<th>WINSTON et al.</th>
<th>IRIS et al.</th>
<th>GERSTL &amp; PРИBBENOW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integral object/component</td>
<td>Functional component</td>
<td>Complex/components</td>
</tr>
<tr>
<td>Collection/member</td>
<td>Collection and members</td>
<td>Collection/elements</td>
</tr>
<tr>
<td>Mass/portion</td>
<td>Spatially included</td>
<td>Mass/quantities</td>
</tr>
</tbody>
</table>

Table 2. (Partial) overlapping of partitive categories in three of the cited taxonomies

related) activities and their ‘parts’. It is expressed by
<includes_phase/is_phase_of>.

Metabolism <includes_phase> Anabolism
Environmental policy <includes_phase> Nature
conservation policy
Transport planning <includes_phase> Road planning

Area-place
It is applied to things that have a spatial extent, indic-
ating the relation between areas and specific places
within them. The latter are inalienable parts of the
whole (areas) in which they are included. However,
like members of a collection, places are not parts be-
cause they functionally contribute to the whole. In
EARTh, it has been applied mostly to geographic
entities and expressed by <spatially_includes/is_spatially_included_in>.

Desert <spatially_includes> Oasis
Earth <spatially_includes> Continent
City <spatially_includes> City centre
Park <spatially_includes> Central park area

Apart from Winston et al.’s proposal, there are also
other taxonomies of the partitive relationship, mostly
developed in the linguistics domain. For example, the
in the work carried out by the Subcommittee on
Subject Relationships/Reference Structures of the
ALA (American Library Association) Subject Ana-
lysis Committee (1997)—who has compiled a master
list of 165 relationships from subject indexing and
cataloguing literature—two main categories are dis-
tinguished: the first, composition partitive relations-
ships, focuses on aggregates or composites of various
members of a class of entities; and the other, whole/
part pairs, is based on structural and spatial rela-
tions and consists of further eight subtypes.

Composition partitive relationships
Whole/part pairs
Non-physical whole/part pairs
Physical whole/part pairs
Arterial whole/part pairs
Anatomical whole/part pairs
Artefact whole/part pairs
Geographic whole/part pairs

Topic inclusion
Discipline/subdiscipline pairs
Whole/attachment pairs
Whole/attachment pairs
Whole/piece pairs
Whole/segmental part pairs

Table 3. Subtypes of the partitive relation from ALA
(simplified version)
Of course, in the framework of ontologies, where attempts to eliminate problems of ambiguity by providing formal definitions of relations are undertaken, the issue of meronymy is greatly discussed, too. An interesting paper dealing with this topic, though in the framework of a broader analysis, is from Smith et al. (2005), who have advanced a Relation Ontology to assist the development of biomedical ontologies, such as the Gene Ontology, and promote their interoperability.

4.2. Some remarks on relation refinement and implementation

Without going further into this analysis, even though the overview is still incomplete, it seems possible to infer an interesting point, that can be applied to all relational patterns. Despite the general agreement regarding a restricted number of basic relationships (namely hierarchical, associative and equivalence), that are in fact used in thesauri and other KOS, a consensus on how to differentiate them into distinct subkinds has still not been—and seems more difficult to be—achieved. Some authors such as Tudhope et al. (2001) have highlighted the risk of an undisciplined extension of the basic semantic model. For this reason, in order to ensure a certain degree of interoperability among advanced systems adopting different solutions, they advocate the adoption of a minimum common denominator—namely the basic thesaural relationships—for different types of applications.

All of this may be partly comprehensible since we are still at an experimental stage in this research field. However, even though, as viewed in the case of the partitive subkinds, there is a more stable consensus among scholars on some more specific relations, the difficulty of univocally determining the ‘final’ set of relations may also be connected to an impossibility of identifying a solution for any circumstance and which could be regarded as equally valid from all viewpoints. Hjørland (2007, 380-381) has underlined, for example, how choices concerning which kinds of semantic relations a system should include have to be related to their practical usage in IR: “In a way, it is the specific ‘information need’ that determines which relations are fruitful and which are not in a given search session. A semantic relation that increases recall and precision in a given search is relevant in that situation.”

The fact is that the further differentiation of the basic semantic relations into subkinds and their better definition do not necessarily guarantee the same results in all applications. Once a shared set is established, this latter may still be dissimilarly implemented. As already said in describing classification, multiple features can, in fact, be ascribed to terms (or objects). Depending on which of these features are made salient in a given context, different relations can be established.

Indeed, the application of the relations in a thesaurus should reflect the knowledge of the subject area that the thesaurus aims to represent (with its paradigms and language games). Besides, it can vary according to different practical concerns and, in any case, to the way in which the criteria defining relations are interpreted and implemented in given circumstances. This might be applied to the partitive relation, too. Depending on all these factors, there could be room left for different ways of conceiving how parts relate to wholes. As underlined in their study of partitive relations by Chaffin & Hermann (1988), even the same pair of objects, and thus of words representing these objects, can be viewed as being connected by different relations once the context changes. This means that, even though cases of strong relational ‘ambiguity’ of such kind are somehow limited to a restricted number, there is not a single way to associate a word-pair to a relation kind (and this concerns also other kinds of relations) (Chaffin & Hermann 1988, 321-22):

The phenomenon of relation ambiguity makes the point that relations are constructed from knowledge of the two concepts related and that a particular relation may make use of some aspects of the two concepts and ignore others …. If two words have more than one relation, then each relation must be based on somewhat different aspects of the two concepts. This point about relation ambiguity may be clarified by comparison with ambiguity in other domains. The closest parallel is with categorization of concepts …. A word pair, more strictly a pair of word senses, may likewise support more than one relation. A relation need not to give equal weight to all aspects of the meaning of the two words. Relations typically emphasize some aspects and ignore others.

An example analyzed by different authors is ‘kitchen-refrigerator’ (Chaffin & Hermann 1988; Iris et al., 1988; Winston et al., 1987). It has been viewed as:
**integral object-component,** when the most important aspect of the refrigerator is considered to be its function in relation to the kitchen (position shared by most of the authors);

**mass-portion,** when the important feature to focus on is size, e.g., in those situations where small kitchens in contrast to large refrigerators are considered (this attribution seems, however, too circumstantial);

**area-place,** when the focus is on the occupied spatial area in relation to the kitchen.

In particular, the possibility of interpreting a word-pair either focusing on the component function and the whole or on the spatial relation occurring between them, pertains, indeed, also to other cases concerning, for example, body structures and geographical items. Remembering that a **component (normally) plays a functional role in relation to an integral object** taken as a whole but is separable from it, and that, instead, a **place is not in this same relation to the area,** but is rather a spatial and inalienable part of it, not always these criteria are easily applicable. A refrigerator normally stands in a kitchen (although it is not an inseparable part of it). From the viewpoint of a kitchen, refrigerators are functional but ‘optional’ parts since it is possible for a kitchen to lack a refrigerator (Cruse 1986). From the point of view of the refrigerator, however, its functional role can be considered apart from its relation with a kitchen (though this is its usual location). Its function, in fact, i.e. ‘to store food (or other products) at a low temperature’ seems more in relation to ‘what’ (to store) than to ‘where.’

This is quite different from the relation, for example, between ‘handle’ and ‘cup’ where the functional role of the handle applies only if it is attached to the cup (of which it constitutes a ‘canonical’ part) and only in relation to that whole. It is interesting also to know that while they regard a refrigerator as being (normally) a functional part of a kitchen, Winston et al. (1987, 433) consider, instead, this latter as “merely a place within a house, not a component of the house” (in other words, ‘house—kitchen’ is an example of **area—place** kind). Yet, this attribution seems to be rather problematic (who would live in a house without a kitchen?).

Summarizing, in our interpretation, neither ‘kitchen-refrigerator’ (where a refrigerator is separable from a kitchen and has a ‘partial’ functional role in relation to it, in the sense that it has a kitchen primarily as its usual functional location), nor pairs like ‘house-kitchen (where the part is not separable from but has a functional role in relation to the whole) seem to fit entirely in one of Winston et al.’s categories and can be, also for this reason, differently classified. This is not only a possible flaw of the taxonomy, but it may also derive from the fact that the complexity of the matter seems to require descriptions based on different perspectives in order to obtain a fuller view. This case seems also to underline the need for more fuzzy-boundaried relational categories: many situations could be more easily classified if conceived as part of a continuum between the two discussed categories.

What has been discussed in this paragraph furnishes, obviously, only some preliminary remarks on this topic. However, to conclude, we may affirm that, while a more elaborated structure can contribute to decrease the level of arbitrariness in the implementation of thesauruses, and this of course is highly recommendable, there is no guarantee that only one valid set of relations exists or that the implementation of more specific relations can provide consistent results in all situations. The hermeneutic principle mentioned in the discussion about classification is, in fact, still relevant, since different choices can be made according to different perspectives and in order to satisfy the needs of different domains and operational contexts.

**5. Conclusion**

A thesaurus is a tool which semantically organizes a domain of knowledge for operational purposes. Its relational semantics is concerned with methods to connect terms with related meanings and designed to support information indexing and retrieval. With focus on hierarchical relations, different aspects of the relational semantics of thesauri as well as the possibility to develop richer structures by differentiating standard relationships into subtypes have been analyzed. We have also examined how semantic issues are implied in thesaurus construction. From a certain viewpoint, a thesaurus relational structure may be regarded as a system providing the representation, for operational purposes, of the meanings of the terms contained in the thesaurus. Thus, theories of semantics, which hold different perspectives about the nature of meaning and how it is represented, affect the way in which the relational semantics of the thesauri is designed.

In traditional approaches to knowledge organization the influence of logical positivism has played a
significant role. And this is also reflected in the current trend towards an increase of formalism and standardization. The search for a more refined relational semantics in thesauri has arisen from this same framework and, according to its advocates, holds the promise to eliminate much of the ambiguity problems.

In our opinion, while it is likely that this field of study will bring valuable results in terms of an improvement of the methodological basis and of a more consistent application, different ways of interpreting meanings and of establishing semantic structures (and thus of organizing knowledge) will continue to be developed, on the basis of different paradigms, domains and operational contexts. Thus, if standardization might be justified in given operational frameworks other solutions should be explored, too. The usefulness of static and monolithic structures is, in fact, rather limited. Tools are, instead, needed that are capable of representing the universe of knowledge domains and structures in its complexity (and also flexible enough to incorporate the continuous changes in languages and meanings, not mentioning how all of this is affected by the development of technology), in order to facilitate access to its constitutive elements (concepts) that are the true object of searching.

Therefore, it is important to consider which contributions may derive from theoretical positions such as those based on hermeneutics and those based on Wittgenstein’s view of language and meaning, which are more inclined to value such complexity (in terms of diversity of perspectives, contexts, rules, etc.). The possibility of their application in thesaurus design and other IR issues has been illustrated, even if this topic needs to be further investigated.

References
Hjørland, Birger Pedersen, Karsten Nissen. 2005. A substantive theory of classification for informa-


Graphic Tools for Knowledge Representation and Informal Problem-Based Learning in Professional Online Communities

Guglielmo Trentin

Istituto Tecnologie Didattiche, Consiglio Nazionale delle Ricerche,
Via De Marini 6, 16149 Genova, Italy, <trentin@itd.cnr.it>


ABSTRACT: The use of graphical representations is very common in information technology and engineering. Although these same tools could be applied effectively in other areas, they are not used because they are hardly known or are completely unheard of. This article aims to discuss the results of the experimentation carried out on graphical approaches to knowledge representation during research, analysis and problem-solving in the health care sector. The experimentation was carried out on conceptual mapping and Petri Nets, developed collaboratively online with the aid of the CMapTool and WoPeD graphic applications. Two distinct professional communities have been involved in the research, both pertaining to the Local Health Units in Tuscany. One community is made up of head physicians and health care managers whilst the other is formed by technical staff from the Department of Nutrition and Food Hygiene. It emerged from the experimentation that concept maps are considered more effective in analyzing knowledge domain related to the problem to be faced (description of what it is). On the other hand, Petri Nets are more effective in studying and formalizing its possible solutions (description of what to do to). For the same reason, those involved in the experimentation have proposed the complementary rather than alternative use of the two knowledge representation methods as a support for professional problem-solving.

1. Introduction

In the discussion group, when trying to best explain one’s viewpoint, oral communication is often accompanied by simple diagrams drawn on the spot either on paper or on a board. One therefore gives a sort of conceptual image (van Lambalgen and Hamm 2001; Stokhof 2002; Wheeler 2006) of the portion of knowledge to be discussed. This in turn triggers a process involving explicit, implicit and tacit knowledge (Polanyi 1975; Nonaka and Takeuchi 1995). The same thing often occurs also during interaction among members of an online professional community. In this case though, instead of paper or boards, \textit{ad hoc} graphic editors are used which allow the online circulation of graphical representations as a support for collaborative interaction. This article, in particular, will refer to two specific methods for the graphical representation of knowledge (Concept Maps and Petri Nets) and related software applications.

2. Graphical Representations

Graphical representations are de facto a language of communication and, like any language, syntactic rules are needed for it to act as a medium in communication between two or more individuals (Donald 1987). Hence, specific graphic languages have been defined
and formalized that are geared towards knowledge representation (hierarchical representations, semantic networks, concept maps, approaches to the representation of procedural knowledge, etc.). Their development has been given considerable impetus from the field of artificial intelligence and, more in general, from all those areas which have attempted “to capture in digital” knowledge domains. They are formally represented so that they can be used by specific software engines: see for example, intelligent systems, decision support systems, semantic webs (Bosch 2006) and simulation systems.

Thanks to their simplicity and effectiveness, some of these graphic languages later spread beyond the specific area from which they originated where their use was often more simplified and less rigorous (Trentin 1991), so that even non-specialists could capitalize on the basic concepts. The question is: when are these graphical representations useful for the professional communities? A first consideration regards their effectiveness in facilitating the multi-perspective study of a given knowledge domain and/or area of exploration: a new knowledge, the solution to a problem, the functionalities of a complex system. The representation of concepts through graphics amplifies, in the eyes of the interlocutors, the existence of multiple interpretations of one subject of study or debate (Cunningham 1991). A second consideration concerns the community’s need for technological aids to improve the flow and organization of community knowledge (Shipman 1993; Prusak 1994; Haldin-Herrgard 2000).

We are aware the knowledge sharing processes (theoretical and procedural) are favored by two types of technological support: one for interpersonal communication and the other for the collection and management of information and knowledge (Auger et al. 2001). Both cases need to give a conceptual schematic representation of the knowledge domain of reference (or portions of it) for a given community. Graphical representations can give an inside view of the conceptual interconnections between elements making up the knowledge that is being discussed and shared. It is therefore an effective way to facilitate the communication of conceptual images as well as the semantic organization of informative, documentary and factual material contained in the community memory (Lave and Wenger 1991). This last aspect is particularly interesting as many research engines now use conceptual representations of the knowledge domain in which they work for the selective recovery of information (for example http://www.webbrain.com).

Before dealing with the experimentation which is the subject of this article, details of the two underlying representation tools of knowledge are summarized here below.

3. Concept Maps

A concept map is a coherent visual logical representation of knowledge on a specific topic which encourages individuals to direct, analyse and expand their analytical skills (Novak and Wåndersee 1991; Halimi 2006). The approach was developed by J.D. Novak (1991) based on Ausubel’s theories (1963; 1968) and Quillam’s studies on semantic networks (1968). Concept maps use diagram representations which highlight meaningful relationships between concepts in the form of propositions, also called semantic units, or units of meaning. A proposition is the statement represented by a relationship connecting two concepts. Therefore, there are two basic features used to construct concept maps: concepts and their relationships (Figure 1).

Besides the two basic features, a concept map is then characterized by hierarchical relationships between concepts and by cross-links between concepts belonging to different domains of the same map.

Various graphic tools for editing concept maps have been developed and the dialogue window in Figure 1 shows of one of the best-known: CMapTool (http://cmap.ihmc.us/). Many of these envi-
environments are able to link the different concepts to a variety of items (documents, images, films, URLs, other concept maps) with the possibility then of converting them into HTML format, thereby creating structured repositories that can be accessed online. This, for example, is one of the possible ways to organize an online community’s shared memory.

Designing concept maps with these software applications is very simple and here, for example, is how one can work with CmapTool:

– after opening a new map and double clicking on the white area, the starting concept may be defined (Figure 2a);
– by clicking and dragging the arrow one can create a link between a new concept and the starting concept (Figure 2b);
– then the two concepts and the relation type linking them have to be described (Figure 2c).

By proceeding in such a way, it is possible to obtain graphical representations like the one reported in Figure 3 showing a maps produced during the experimentation described here.

When very complex knowledge domains have to be described, such as the Clinical Audit in Figure 3, the corresponding concept maps tend to become much larger and difficult to manage. For this reason, CMapTools provide a function to compress/expplode sections of the map being drawn. For example, by clicking on the symbol “>>” that appears to the right of “evidence-based practice”, the map linked to that concept expands (see Figure 4). Then clicking on the symbol “<<” will take you back to Figure 3.

4. Petri Nets and Procedural Knowledge Representation

Petri Nets provide an effective way to describe and analyze models, whether complex systems, processes, knowledge domains, etc. (Peterson 1981). On account of this characteristic, they are often used in the graphical representation of procedural knowledge.

4.1. Resources and activities

A Petri Net is an oriented graphic in which two node types are represented: resources (indicated with circles in Figure 5) and activities (indicated with segments)—in literature on Petri Nets these nodes are respectively called places and transitions (Peterson 1981).

A graphic arc that is directed from a resource to an activity indicates that the resource is necessary to carry out that activity. Similarly an arc that is directed from an activity to a resource indicates that the resource is the product of the same activity.
Figure 3. A concept map on the Clinical Audit developed with CMapTool

Figure 4. Example of a complex concept expansion
What has just been listed are, so to speak, the basic “ingredients” to give shape to Petri Nets according to the use suggested within the experimentation referred to here. In actual fact, the theory presupposed by the Petri Nets is much more articulated and rigorous (Peterson 1981). In our case only the key concepts have been used to enable the two communities involved to assess the general philosophy governing the specific approach.

Starting from an initial Petri Net - in attempting to describe the process/procedure or knowledge domain with even greater precision - activities, resources and links are often increasingly added. This therefore produces very complex graphs that are hard to process and read. A good method to overcome this drawback is to describe the network through successive refinements (or stages), expanding it using a top-down approach (Trentin 1991). In the first stage an overall (undetailed) representation is given of what one wants to describe. The resources and main activities are reported together with their respective interconnections. In the same network the complex activities are then highlighted that will be described in more refined detail in a specific sub-network. See, in Figure 6, activity “AC development” represented with a grey square.

The following stage involves developing the refinement sub-networks giving a detailed description of the more complex activities. For example, Figure 7 reports the refinement of activity “AC development” shown in the Petri Net of Figure 6.

The refinement process is iterated until the desired level of detail given to the representation is attained.

The refinement activity is a consequence of the need to foster the so-called “functional abstraction” (Stein 2002), the process through which the attention of the individual or whole group/community focuses on one aspect of what is being described at a time.

This is a process developed stepwise. It begins with an overview of the subject matter, such as a professional issue, where the key elements characterizing it are identified (macro-representation of the domain). In the following steps, each key element is isolated and described in more detail by breaking it down into less complex sub-elements (for example, a complex activity is broken down into sub-activities). This is
done by trying to abstract as much as possible from what is within the confines of the element that is considered one by one (the other elements), to guarantee maximum success of its specific analysis.

Should this refinement step be inadequate for a deep analysis of the element being dealt with, the refinement process is iterated until the level of detail is considered the most functional to reach the final objective (analyzing a situation, solving a problem, describing a complex system).

5. Research Issue

The use of graphical representations is very popular in information technology and engineering. Although the same tools could be applied effectively in other areas, they are not though since they are not well known or are completely unheard of. This is due to study curricula and/or training courses where there is no occasion to learn these techniques and technologies since they are not considered important for a given disciplinary/professional area.

This is the reason why - within the two specific projects aimed at fostering the launch and development of professional communities in the health care sector - research was carried out on the use of graphical approaches to professional knowledge representation. The aim was to analyze and discuss their actual usability and effectiveness in fostering collaborative interaction, debate and reciprocal clarification during a process geared towards examining a specific professional theme/issue.

6. Experimental Setting

Two distinct professional communities have been involved in the research. The first (Audit community) was made up 31 head physicians and health care managers pertaining to Local Health Unit 11 of Livorno (Tuscany Region) who had the task of dealing with the theme of Clinical Audit, the key elements characterizing it and the working methods to carry it out. The second (Alert community) formed by 18 technical staff from the Department of Nutri-
tion and Food Hygiene coming from all the health care units in Tuscany. In their case, the task was to define the organization of a Regional Working Group on the problem of managing food alerts.

In both cases, as already mentioned, concept maps and Petri Nets have been proposed as methods for graphical representations of knowledge. The development of each graphical representation has been divided into three stages:

- a face-to-face meeting for the first familiarization with the graphic approach and the related editing software;
- two weeks of online collaborative activities in sub-groups;
- a closing meeting to evaluate and compare the graphical representations produced, and to discuss the online collaborative process implemented to produce them.

The participants were divided into sub-groups of 5-6 units and were asked to structure their work into two one-week periods:

- individual drawing up of one’s draft of the graphical representation;
- sharing of graphical representation and convergence towards one single sub-group version of it.

To co-construct the two representations the following applications have been used:

- CMapTool (http://cmap.ihmc.us/) and WoPeD (Workflow Petri Net Designer) (http://www.woped.org/) respectively for the development of concept maps and Petri Nets;
- Moodle (http://moodle.org/) as environment to run interpersonal group communication.

7. Methodology

At the end of the collaborative activity, the participants were given a questionnaire divided into 4 sections:

A. Learnability, intended to pinpoint the times and possible learning difficulties of the approaches to the formal representation of knowledge used in the experimentation.

B. Study and/or problem-solving, intended to research the perception of the general usefulness of the tools proposed for the study activities, analysis and search for solutions.

C. Usefulness on an individual level in one’s own professional practice, intended to research the perceived usefulness of tools proposed in relation to an individual use in one’s own professional practice.

D. Usefulness in facilitating collaborative group work, intended to discover the perceived usefulness of tools proposed in fostering or not fostering group work when dealing with aspects related to their own professional practice.

In the questionnaire, two questions are associated with each survey indicator: one with a closed-ended answer based on attributing a score (on the Likert 1-5 scale); the other with an open-ended answer asking to explain the attribution of the above-mentioned score or to give further information about the same indicator. 25 participants belonging to the Audit community and 16 to the Alert community answered the questionnaire anonymously.

8. Results

The survey data revealed positive evaluations regarding the professional use of proposed graphic formalization methods. However, there were various and sometime considerable differences between what was expressed by the two communities. This likely to be related to the different roles covered by the respective individuals: on the one hand, positive but lower scores were given by the Audit community made up mainly of people with a managerial role; on the other hand, higher scores were assigned by the Alert community made up of staff with a more technical role. A more analytical examination of the participants’ answers is provided in the next section.

8.1. Learnability

As shown by Table 1, both groups stated that they found it more difficult to enter the logic of the Petri Nets than the concept maps.

<table>
<thead>
<tr>
<th>Learnability</th>
<th>Audit</th>
<th>Alert</th>
</tr>
</thead>
<tbody>
<tr>
<td>How easy has it been for you to master the logic and syntax of the concept maps?</td>
<td>3,1</td>
<td>3,7</td>
</tr>
<tr>
<td>How easy has it been for you to master the logic and syntax of the Petri Nets?</td>
<td>2,6</td>
<td>2,8</td>
</tr>
</tbody>
</table>

Table 1. Average data relating to answers on learnability
It is a fairly common reaction, met in other similar experimentations (Trentin 1991; Stein 2002), and should be related to the greater effort of abstraction (and of dissection) that the top-down development of a Petri Net requires. The free answers given by the participants show how the use of concept maps seems to best mirror their way of coping with professional problems i.e. considering the elements characterizing them all together and simultaneously. The use of the Petri Nets, with a top-down approach, generally baffles the professional not used to functional abstraction mechanisms which are more familiar in information technology and engineering. This was confirmed by directly observing the participants’ first approach towards elaborating a Petri Net where individuals tended to draw a very detailed, and therefore complex graph already at the overview stage of the knowledge domain. Some open answers given by participants pointed out, among the probable causes of difficulties, how they are used to a sequential approach to analyzing problems which is closer to the logic of flow-charts (used occasionally by some of them) than to the logic of top-down.

8.2. General usefulness for study activities, analysis and problem-solving

To best understand the convergences and divergences expressed by the participants on this point, we will firstly make a quantitative comparison of the average scores assigned by the two communities and then summaries the usefulness of the two approaches in relation to every single activity indicated in the questionnaire.

8.2.1. Quantitative comparison of the scores assigned by the two communities

As can be observed in Figure 8, the trends of average scores attributed by the two communities are fairly similar even though they are quantitatively different. The only divergence that is rather noticeable corresponds to the use of concept maps for study activities. In this regard, 8 members of the Audit community justified the low score claiming that drawing up a concept map on a given topic can be done only if one already has sufficient knowledge about it. They

![Figure 8. Quantitative comparison between the average scores assigned by the two communities in relation to the usefulness of graphical representations in their profession](image-url)
therefore think that the use of the concept maps can be more useful as a self-check tool of one’s learning than as an aid to studying (at least the basics). On the other hand, the rather high score attributed by the Alert community should be related to their idea of using the concept maps as a tool to support the collaborative study processes.

8.2.2. Summary on the different usefulness of the two approaches

Apart from the deviation between the quantitative evaluations formulated by the two groups and the above-described divergence, from the graph in Figure 8 it can be deduced that:

- the graphical representations are considered useful particularly for analysis and problem-solving activities and less useful for study activities. The evaluation of the Alert Community is an exception to this in correspondence with the use of concept maps;
- both communities showed concordance (despite attributing rather different average scores) in evaluating that the use of the concept maps are more recommended in analysis activities whilst that of the Petri Nets in problem-solving activities.

To sum up, the participants indicate that the concept maps are more useful in describing “what it is” whilst the Petri Nets in describing “what to do to.”

8.3. Usefulness of graphical representations on a personal and group level

After the general considerations, described in the previous sections, participants were asked to evaluate the perceived usefulness of the two graphic methodologies as a tool for both personal and group use in their professional practice. Here are their evaluations:

<table>
<thead>
<tr>
<th>Personal usefulness of graphical representations</th>
<th>Audit</th>
<th>Alert</th>
</tr>
</thead>
<tbody>
<tr>
<td>How much do you think Concept Maps can/could be useful in your professional practice?</td>
<td>3,3</td>
<td>3,8</td>
</tr>
<tr>
<td>How much do you think Petri Nets can/could be useful in your professional practice, to describe complex situations/systems?</td>
<td>3,2</td>
<td>3,6</td>
</tr>
</tbody>
</table>

Table 2. Average data relating to the personal usefulness of graphical representations

As can be seen, both communities gave between average and high average scores regarding the personal usefulness of graphical representations.

The attitude changes when instead the same tools are considered for collaborative group activities.

<table>
<thead>
<tr>
<th>Usefulness of graphical representations in group work</th>
<th>Audit</th>
<th>Alert</th>
</tr>
</thead>
<tbody>
<tr>
<td>How much do you think Concept Maps can/could be useful in group work?</td>
<td>3,7</td>
<td>4,1</td>
</tr>
<tr>
<td>How much do you think Petri Nets can/could be useful in group work, for the representation of procedural knowledge?</td>
<td>3,8</td>
<td>3,8</td>
</tr>
<tr>
<td>How much do you think Petri Nets can/could be useful in group work, to describe complex situations/systems?</td>
<td>3,7</td>
<td>3,9</td>
</tr>
</tbody>
</table>

Table 3. Average data relating to the usefulness of graphical representations in group work

A comparison between Table 2 and Table 3 shows how the participants underline how graphical representations are more useful in group work than in individual work. Here, both communities have shown a certain convergence of opinion, although there are the usual deviations in average values.

From the diagram in Figure 9 it is interesting to observe how there is an appreciable divergence between the two communities regarding the usefulness of the Petri Nets. The Audit community believe they are more effective for representation activities of procedural knowledge. On the other hand the Alert community consider them more useful for those activities connected to the description/analysis of complex systems. This is for both individual and group activities. Again, the divergence of opinion is likely to be related to the members’ role within the two different communities in the respective local health units.
9. Conclusions

Perhaps the most interesting result emerging from the research is the idea of combining the use of the two graphic tools for professional problem-solving activities. In particular, as the participants indicate explicitly in some answers, the concept maps are believed to be more effective in analyzing the knowledge domain related to the problem to be faced. On the other hand, the Petri Nets are thought to be more effective in studying and describing the procedures to solve the very problem.

Indeed this is confirmed by the typical stages characterizing problem-solving strategies (Heller and Reif 1984; Gick 1986):

1. analysis of reference scenario related to the problem;
2. description of what is already known regarding the specific problem;
3. formalization of the problem and of its possible breakdown into sub-problems;
4. identification of actions to undertake to provide a solution to the problem and/or individual sub-problems where it can be broken down;
5. identification of necessary resources to carry out actions determined in the previous point

As can be observed, in the high stages (see points 1-2), where the question is to define the problem in terms of “what is it”, the concept map would in fact appear to be the most suitable tool. In the successive stages (3-4-5), the Petri Nets would instead have the advantage of favoring the procedural description of “what to do to”, at a macro level (solution overview) as well as micro level (solution details to sub-problems comprising the general problem).

With regard to the procedural representation of knowledge, it is worth pointing out how some participants found Petri Nets more effective than flow-charts in describing processes/solutions. This is due to at least two reasons:

- because besides indicating the link between activities characterizing a process, Petri Nets require the necessary resources for their development to be defined (flow-charts focus only on the statements);
- the top-down refinement helps focus step by step on the specific parts of the process and therefore

Figure 9. Comparison between the average scores assigned by the two groups regarding the usefulness of graphical representations respectively for individual and collaborative use
avoids managing the complexity of what is being studied/analysed with just one graphical representation.

These are a fairly interesting conclusions that could lead to new developments in researching technological solutions to support the integration of the two methods of formal knowledge representation discussed here. The solutions need to be able to offer, through the same software environment, support functions to the conceptualization and to the proceduralization in problem-solving activities.

These activities, as is known, provide the ideal opportunity to trigger informal peer-to-peer learning processes which are typical in online professional communities.

References


The Immediate Prospects for the Application of Ontologies in Digital Libraries

Jody L. DeRidder

Digital Library Center, James D. Hoskins Library, University of Tennessee, Knoxville, Tennessee. USA, <jo@utk.edu>

Jody L. DeRidder received her M.S. in Computer Science from the University of Tennessee in 2002, after developing repositories for the Open Archives Initiative in its alpha phases. As the lead developer for the Digital Library Center of the University of Tennessee Libraries, she has built, customized and altered software to create interoperable digital library systems which provide usability features beyond the norm. Nearing completion of her M.S. in Information Sciences, her research interests have turned to interoperability between systems to support usability, sustainability in digital libraries, and the application and use of ontologies via automated cross-mapping by query engines.

ABSTRACT: The purpose, scope, usage, methodology, cross-mapping and encoding of ontologies is summarized. A snapshot of current research and development includes available tools, ontologies, and query engines, with their applications. Benefits, problems, and costs are discussed, and the feasibility and usefulness of ontologies is weighed with respect to potential and current digital library arenas. The author concludes that ontology application potentially has a huge impact within knowledge management, enterprise integration, e-commerce, and possibly education. Outside of heavily funded domains, feasibility depends on assessment of various evolving factors, including the current tools and systems, level of adoption in the field, time and expertise available, and cost barriers.

1. Introduction: defining ontology

Each of us has a slightly different way of looking at the world. Across cultures and research areas, these differences become palpable. What is clearly understood within a community may be unknown elsewhere and technically specific terminology needs to be translated, as if to a different language, for the general user. For applications to be able to serve us in search and retrieval across all these variations, human knowledge needs to be made comprehensible to computer programs. Building an ontology requires capturing concepts (including implicit ones), the relationships between them, and any constraints on those relationships (de Bruijn 2003, 35). In technical terms, an ontology represents a “language” of concepts, relations, instances and axioms (de Bruijn and Polleres 2004), which enable computer applications to logically reason out solutions or adapt queries. Stanford University offers a sample ontology application which suggests wine selections for your choice of food includes encoding examples and explanations (Hsu 2003). To illustrate an ontology description of an object, a graphic example of an ontology application to an audio tape of a performance of a single concerto (in the ABC ontology) is shown in figure 1 (Hunter, 2001).

1.1 Points to consider

For ontologies to be useful and feasible in digital libraries, several requirements must be met. First, there must be evidence that they are helpful to users. Usefulness must outweigh the cost and effort of creation and maintenance. Here we must consider further the identification of our user audiences, and the purpose and scope of what we wish to accomplish. Secondly, what is the state of the art? What parts of this territory have been mapped out, and what are still murky waters? Is there, or will there soon be, broad support for the use of cross-mapped ontologies? If the road is clear and support is avail-
able, it behooves us to make our digital libraries accessible via ontology mapping, to increase accessibility, interoperability, and to leverage the work in the broader arena to meet our constituents’ needs. If it will be years before the path is paved, standards will likely change rapidly over that time. Those with the funding and the capability can lead the way, contributing to the development of standards and interoperability. If funding and capabilities are limited, it is wiser to wait till the paths are well-laid, and the process is easier. Thirdly, we need user-friendly tools and methodology. What are the steps? What personnel and tools are needed? As the field is still clearly in the beginning stages, an overview of current research and development is provided for further investigation. Finally, we must seriously consider the costs. What level of funds, personnel, and expertise are available?

1.2 Benefits

As systems grow in decentralized manners, semantic heterogeneity is inevitable; how do we provide functional search and retrieval across distributed digital libraries? Searching by keyword retrieves irrelevant information when a term has multiple meanings; and information is missed when multiple terms have the same meaning. In addition, concepts that may not be represented by the terminology in the document or metadata are not available to searchers. Information retrieval is a negotiation process, and as digital content multiplies, users need assistance in wading through the results of their searches. A comparison of precision and recall between full text searching, latent semantic indexing, and ontology-based retrieval (with manual assignment of concepts to query) finds ontologies capable of providing far better retrieval efficiency (Paralić and Kostial 2003).

Digital libraries routinely provide their services without human assistance; thus it is essential that their metadata be suitable for computation, supporting inference. The reference interview is not available; therefore, computer applications need to be able to reason about their contents to reformulate queries, deduce relations between works, and customize services to the task and user. This is only possible via ontologies (Weinstein and Birmingham 1998).

Imagine a user entering a query, and the computer application offers different meanings for the entered terms; the user selects the intended meaning, or chooses one of the related terms offered. The query engine transforms the query into a language that matches the terminology used in describing the data sources. In addition, it locates material related to your query, based on logical deduction and inference, offering these results on the side. In this manner, relevance and pertinence are improved, and browsing is enabled. With ontologies, we enable computer applications to perform intelligent searching instead of keyword matching, query answering instead of information retrieval, and to provide customized views of materials. A standardized vocabulary referring to natural language semantics enables automatic and human agents to share information and interoperate functionally (Fensel et al.2003c).
1.3 Depth and breadth

There are many different ways to classify ontologies; two of the most useful reflect the depth and the breadth of the ontology. In the depth dimension, the specificity of the ontology determines its “weight.” Lightweight ontologies are little more than taxonomies, and include only concepts and their properties, relationships between those concepts, and controlled vocabularies. Heavyweight ontologies also include axioms and constraints that increase the capability of a computer application to logically reason with the data given. Dublin Core might be considered an extremely light-weight ontology, whereas Cyc (created using the Knowledge Interchange Format, a proposed standard) may be the most extensive top-level ontology currently in existence (de Bruijn 2003, 6-9). (Two limited open-source versions of this encyclopedic ontology are available: OpenCyc and Research Cyc.) In the breadth dimension, there are general (top-level, or global) ontologies, domain ontologies (specific to a particular area) and application ontologies, which describe concepts depending on the task as well as the domain (some refer to application ontologies as another form of domain).

1.4. Cross-mapping issues

In order to provide searching via natural vocabulary, a mapping is needed from the natural language of each user group to the entries in each metadata vocabulary. This is known as an “entry vocabulary index” or EVI. In addition, to search across databases, it is necessary to have mappings between each possible pair of system vocabularies, or ontologies. Mapping between ontologies must be done by people competent in both domains; the current status is that human assistance in mapping will likely be necessary for some time to come, for high quality mappings (Bockting 2005).

Problems in cross-mappings can be of several types. Data objects of the same name may describe different real-world elements; concepts may be ascribed to different levels of the metadata structures (an attribute in one ontology may be a class in another); conceptual approaches may preclude a functional correspondence; descriptions of a single real-world element may vary considerably and conflict with one another; and one of the ontologies may have incorrect information (Adam, Atluri, and Adiwijaya 2000). A concept in one ontology may not exist in another, or may have an entirely different meaning. For example, in the Harmony Project, members of the closely-related domains of digital libraries and cultural heritage and museum communities sought to merge the digital library ABC Ontology (Lagoze and Hunter 2001), with the CIDOC (International Committee for Documentation of the International Council of Museums) Conceptual Reference Model (ICOM/CIDOC and CIDOC CRM 2005). They uncovered cultural biases particularly in terms of the nature of change; while both ontologies were concerned with change over time, one modeled the change of objects, while the other modeled changes in the context and meaning for those objects (Doerr, Hunter, and Lagoze 2003). A comprehensive overview of the problem areas of mapping, including variation of expressiveness and the differing modeling paradigms or styles, is discussed by Klein, and diagrammed in figure 2 (Klein, 2001).

![Figure 2.](image-url)
An IMLS-funded effort (National Leadership Grant No. 178), based on prior research partially supported by a DARPA (Defense Advanced Research Projects Agency) contract, explored the feasibility of cross-mapping vocabularies of numeric data sets and text files (Buckland et al. 2007). It was discovered that the vocabularies for topical categorization vary greatly, requiring interpretive mappings between systems, and that specification of geographical area and time period are problematic. Both names of places and of time periods are culturally based, unstable, and ambiguous. The use of geospatial coordinates is suggested as the only effective method of relating locations to search terms, which means that both gazetteers and map visualizations become critical to implement search retrieval in a user-friendly manner. A similar application needs to be developed for time periods, and this issue is being addressed in a subsequent IMLS-funded study by the Electronic Cultural Atlas Initiative (Electronic Cultural Atlas Initiative 2006). Among other objectives, the intent is to contextualize objects in library and museum collections by using or adapting existing and emerging standards and protocols. This initiative is described further in (Petras et al. 2006).

Ontologies must be expected to evolve over time as knowledge and understanding grow, and terminology changes. Their mappings to other ontologies must also evolve, and this evolution may require change in other ontologies to which they are mapped (de Bruijn and Polleres 2004, 11). Thus the initial effort to develop ontologies is insufficient; they must not only be maintained but also versioned over time, and compatibility with other ontologies considered with each evolution. Cross-mappings are rare, expensive, time-consuming, and difficult to maintain. With 135 semantic types and 54 relationships, the Unified Medical Language System Metathesaurus is a notable example (Smith et al. 2004).

1.5. A bird’s eye view

It is insufficient to consider ontology mapping as a singular or only a local problem. Many differing ontologies already exist with overlapping domains of knowledge and application (de Bruijn 2003). And there are at least three basic conceptual approaches to interoperability: a global ontology to which all local ontologies are mapped, a peer-to-peer system (where mappings exist between local ontologies where needed), and a combination of the two. A central, heavyweight global ontology is clearly preferable for computer applications, as one-to-one mappings of all involved ontologies does not scale. However, obtaining global agreement on controlled terms and relationships is infeasible, so a layering approach based on generality is more likely to succeed, with mapping between domains and higher level ontologies as needed (Meersman 1999). A single general light-weight ontology to be shared by multiple domains was explored by (Stuckenschmidt and van Harmelen 2005). After developing their framework, the authors stated that the shared ontology can only be developed if all sources of information are known, and the conceptualization of each source is accessible; they concluded this was only feasible for a single domain (Stuckenschmidt and van Harmelen 2005, 249). De Bruijn and Polleres add that a limitation to this approach would be the likely lack of agreement on the interpretation of the concepts in the shared ontology by all the authors of local ontologies (de Bruijn and Polleres 2004).

Another possible middle ground between the peer-to-peer approach and the central core ontology method, would be to implement layers or a hierarchical application (de Bruijn and Polleres 2004). One way to envision this is to compare a scientific discipline with a group of islands, where each area of research is an island, and each island has a further breakdown of specificity into “dialects.” If a single island had 3 dialects, each dialect would be a Level 1 ontology, probably the most specific in terminology. A shared ontology for the entire island would be a Level 2 ontology. Islands (or domains) could map to one another as needed. A shared ontology for the group of islands would be a Level 3 ontology, the most general so far. Other sets of islands could have similar structure, and again, the hierarchy could continue as needed, but with a distributed, organically growing base rather than a single top-down application. This may be the only feasible solution, as it reflects the grassroots approach and grows as needed.

2. State of the art

Currently, the semantic search engine Swoogle states that there are at least 10,000 ontologies in use on the WWW, and provides a list of ontology repositories, semantic web search engines and crawlers. The 2005 version of Swoogle indexed 337,182 documents, while the 2006 version currently lists their number of documents at 2,030,039 (Swoogle 2006), a major increase. This cursory comparison indicates a growing interest in the implementation of ontologies.
2.1 Within domains

Ontologies seem to have already found a home in instructional technology, as an outgrowth of KOS (Knowledge Organization Systems). The primary difference is that ontologies apply logic to the relationships (Binding and Tudhope 2004). Other differences are that existing KOS lack conceptual abstractions, semantic coverage, consistency, and automatable processing (Soergel et al. 2004). Ontologies are important to education because concepts and the relationships between them "provide a powerful, and perhaps the only, level of granularity with which to support effective access and learning" (Smith et al., 2004, 2). A portal already exists for sharing tools, projects, research and information for ontology use in education (Dicheva et al. 2006), and a commercial success in the education arena is Xyleme, which depends upon the existing heterogeneous XML structure in documents for pattern-matching, mapping, encoding, and creating "views" for abstract query response (Aquilera et al. 2000).

The Alexandria Digital Earth Prototype (ADEPT), currently in use for teaching geography courses at the University of California, employs an ontology to link the current lecture material to a graph showing its relation to other concepts, and also links to examples from the digital library. All three views are presented at the same time, to give students the context and examples they need to make sense of what the teacher is trying to communicate. In addition, the ontology supports a Virtual Learning Environment that lets the teacher create, use, and re-use learning materials in different fields of science and in various learning environments (Smith et al. 2004). Yet here the content of the digital library itself is limited to examples, primarily images and graphs. For digital libraries containing complex materials, there exists the need for two levels of access: discovery of resources, and discovery within the resources, the latter of which requires the creation of descriptions of semantic and internal structural organization through resource decomposition. The GREEN digital library project explored the problems and possibilities in this area, using term extraction algorithms, performing text analysis, and extending a combination of metadata schemes (LOM for learning objects and MatML for materials). This group noted the need for a convergence of metadata schemes and robust mechanisms for navigating a complex associational web of resources (Shreve and Zeng 2003). Clearly, the ability to locate specific content, regardless of its location within materials, would be extremely useful for isolating information and minimizing the time spent sifting through search results. As the quantity of materials online explodes, findability becomes critical.

An example of ontology use in enterprise integration would be the Unified Medical Language System (UMLS), which provides services for computer applications across a multitude of health-industry areas. The UMLS Metathesaurus is a compendium and synthesis of more than 100 different thesauri, classifications and code sets for health care, billing, statistics, medical literature, research and resources, and requires constant updating and renovation. The Metathesaurus preserves the many views present in the source vocabularies, as each may be useful for different tasks. Hence, it must be customized to be effective in any one application (U.S. National Library of Medicine, March 2006a). UMLS includes a Semantic Network to "provide a consistent categorization of all concepts represented in the UMLS Metathesaurus and to provide a set of useful relationships between these concepts" (U.S. National Library of Medicine March 2006b). In addition, the SPECIALIST Lexicon provides a general English vocabulary that includes biomedical terms, for Natural Language Processing (NLP), to improve searchability for the general user (U.S. National Library of Medicine, March 2006c).

E-Commerce potential is clearly indicated in the level to which ontologies have already proven their value in critical government defense, finance, and manufacturing. An example in the business arena is Australia’s InfoMaster. In the United States, Ontology Works, founded in 1998 by former members of the intelligence community, currently serves the critical needs of such clients as the U.S. Department of Defense, the U.S. Department of Justice, Science Applications International Corporation, Boeing, Northrop Grumman, and the Sierra Nevada Corporation. Ontology Works is a highly successful commercial venture, and claims to have the most sophisticated ontology-driven database on the market (Ontology Works, 2005). Another commercial success is Ontoprise, a deductive, object-oriented database system, now available via Ontoprise.

MOMIS (Mediator environment for Multiple Information Sources) has been used to model a tourism information provider system. In the MOMIS Integration Methodology, local source schemata are extracted. If the source material is unstructured, text is extracted, analyzed, and an XML schema is generated. Then a meaning for each element of the source schema is chosen from a lexical database of English,
WordNet (prompts for choices are given to a human; the choice is manual). A common thesaurus, a global schema, and sets of mappings to local schemata are generated. Finally, a meaning is assigned (semi-automatically) to each element of the global schema. The query manager then rewrites the incoming global query as an equivalent set of queries to match the local source schemata; local sources are queried with these, and the resulting responses are fused and reconciled into a final response (Bergamaschi et al. 2005).

Exploration has been made into non-textual content as well. Annotation of historical images with a domain-specific ontology enables users to retrieve images for which they inadequate historical knowledge and keywords (Soo et al. 2002). An Amsterdam research group has developed a Visual Ontology Using MPEG-7 and WordNet, which supports descriptions of colors and shapes of objects, to support automatic annotation (Hollink et al. 2005). By extracting and analyzing visual features, mapping clusters of sequences and patterns to ontological concepts, another experiment has demonstrated the feasibility of semi-automated ontology annotation of domain-specific videos (Bertini, et al. 2005). In a fourth model, audio tapes of sports broadcasts were annotated (Khan, McLeod, and Hovy 2003), though the text analyzed was extracted from the closed captions that came with the audio objects. In this project, only three relations were modeled (isA, Instance-Of, and Part-Of), and an automatic query expansion mechanism was built using WordNet as a generic ontology, though they found it too incomplete to functionally model the domain.

According to (Ontology Works 2005), the leading research groups in ontologies are IFOMIS (The Institute for Formal Ontology and Medical Information Science), ECOR (European Centre for Ontological Research), LOA (Laboratory for Applied Ontology), and NCOR (National Center for Ontological Research). Based on the number of recent ontological projects, Stanford University’s Knowledge Systems, Artificial Intelligence Laboratory and the Sirma Group’s OntoText Semantic Technology Lab should perhaps be added to this list.

2.2. Across domains

One of the primary purposes of cross-mapping is to allow searching of heterogenous resources from a single interface. The Digital Government Research Center Energy Data Collection project used an overarching ontology (SENSUS) to provide searching across over 50,000 database tables, manually defining the domain model with 500 concept nodes, then mapping them with intentionally vague semantic meaning to the possible 70,000 nodes of the larger ontology. While much of the model building was automated, it was far from simple to create a coherent domain model out of the variation of metadata and domain terms within the databases. The end product cannot support automated inference, but does enable browsing and non-expert searching with familiar terms (Hovy, 2003).

OntoMedia, an opensource effort, builds on the CIDOC Conceptual Reference Model and the IFLA-NET (International Federation of Library Associations and Institutions) FRBR model (Functional Requirements for Bibliographic Records) to facilitate the annotation of semantic content of multimedia. It provides the user with a graphical user interface with metadata indexing and search capabilities, for organizing multimedia collections, though the ontology is presented as a general, high-level ontology for reuse across domains (Lawrence et al. 2005).

Semantic Interoperability of Metadata and Information in unLike Environments (SIMILE) is a joint project of MIT Libraries and MIT Computer Science and Artificial Intelligence Laboratory, which leverages and extends DSpace. The intent is to enhance general interoperability across distributed information stores of varying types, and to provide useful end-user services for mining that material (Leuf 2006, 223-4). In an early prototype of the project, VRA Core (Visual Resources Association Data Standards Committee) and IMS LOM (Learning Object Metadata) were translated into RDF schemas with enrichment obtained from Wikipedia and the prototype OCLC Library of Congress Name Authority Service. Then the datasets were transformed from XML to RDF/ XML using XSLT. While the developers were able to automate linkage of RDF datasets using string similarity techniques, the approach was error prone and results had to be manually reviewed. In addition, the enrichment techniques could be automated as well, but again, required human intervention to verify the validity of the data produced (Butler et al. 2004).

3. Fundamentals

3.1. Methodology

A recent analysis of the state of ontology engineering bemoans a lack of guidance, unified methodology,
cost benefit analysis tools, and selection support to choose engineering approaches (Simperl and Tempich 2006). Real-world applications require comprehension of the scope and progression of the project, customizable workflows, user-friendly tools, and automation of the majority of the tasks. While several ontology management tools are relatively mature, many necessary ontology engineering activities are not yet adequately supported by technology, and critical aspects, such as automation of ontology creation, application, and mapping are still being researched. The basic model for implementation of an ontology (without consideration of ontology mapping; see figure 3) includes a feasibility study, domain analysis, conceptualization, encoding, maintenance and use (Simperl and Tempich 2006).

One unusual investigation tested the hypothesis that the more indexing is geared toward the user task, the better the results. Kabel, Hoog, Wielinga and Anjewierden (Kabel et al. 2004) compared the efficiency, effectiveness, precision of use, and quality of results when users were given access to keywords versus a domain index versus an instructional index, for creating lesson plans. The domain index was content-based, with specific terminology. The instructional index provided classification of objects by use in instructional material, and hence was task-oriented (an application ontology). An example of this would be a “behavioral description” with “specific” scope, and the instructional role of “illustration.” Their hypothesis was generally correct. The domain index provided more efficient, effective search and retrieval than the keyword search, and the instructional index provided better precision than the use of keywords and domain indexing. Hence, it appears that we need to clearly understand the needs of our users, in order to choose the type of ontology that will actually provide the specificity they need for the task at hand.

ScholOnto (Shum et al. 2000), for example, is an effort to develop an ontology for discourse about research, rather than for the research itself, which is an

3.2. Purpose and scope

Before choosing, adapting, or creating an ontology, the purpose and the user audience must be determined. If the domain is clearly delineated and there is no desire for interoperability or cross-mapping to outside ontologies, the scope and direction are simplified. If, however, the desired outcome is more diverse and interoperable, the choices made in this assessment will be both critical and complex.

Figure 3. Ontology Engineering Activities
interesting twist. Designed to provide an ontology for scholars to interpret, discuss, analyze and debate about existing literature, ScholOnto (developed using OCML (Operational Conceptual Modeling Language) overlays existing metadata and does not attempt to directly describe the content of the research. Instead, the ontology provides a structure to clarify the intellectual lineage of ideas, their impact, scholarly perspectives on those ideas, inconsistencies in approaches or claims, and convergences of different streams of research (Shum et al. 2000, 3). Here, the comments about the literature become the objects for retrieval and for building new structures to define the usefulness of the object. This is a social networking function, an interactive community-created layer over the research itself. This could be an invaluable way to add context and clarity to understanding and exploration of a domain. Thus, the application of ontologies to digital libraries might not be in querying the documents themselves, but in building relationships and connections and social context around the documents.

3.3. Conceptualization

If ontologies exist that can be adapted to the purpose at hand, tools are needed to perform such adaptation. If an appropriate ontology does not yet exist, tools are needed for modeling and constructing the ontology. Selecting or creating an ontology involves a fundamental tradeoff between the degree of complexity and generality versus the degree of efficiency of interpretation and reasoning within the language (Weinstein and Birmingham 1998, 35). Maximum consideration must be given to the desired services. The following findings are intended to provide a starting point for further exploration.

One possibility is that of creating ontologies out of existing metadata schemes or thesauri, adapting and adding as needed. The more complex and structurally coherent the metadata scheme, the more feasible this may be. One effort under development is an adaptation of the AGROVOC Thesaurus, developed and maintained by the Food and Agriculture Organization of the United Nations (Soergel et al. 2004). An older effort to transform MARC (Machine Readable Cataloging) uncovered difficulties in the varying dimensions and multiple levels of granularity containing partial descriptions, which is a requisite feature of bibliographic data (Weinstein and Birmingham 1998). Another possibility is creating an ontology from scratch, using existing models to pave the way. OCML (Operational Conceptual Modeling Language) supports the construction of ontologies and problem solving methods, and is supported by a large library of reusable models (via the WebOnto editor). Currently in use by several projects, OCML is available free of charge for non-commercial use.

Building on previous work is a third option, and the one which offers the greatest variety of tools at present. Many of these are domain-specific.

The ABC Metadata Model Constructor fundamental classes for digital libraries were determined by analyzing commonalities between Dublin Core, INDECS (Interoperability of Data in e-Commerce Systems), MPEG-7 (Multimedia Content Description Interface), CIDOC (International Committee for Documentation of the International Council of Museums) Conceptual Reference Model and the IFLANET (International Federation of Library Associations and Institutions) FRBR model (Functional Requirements for Bibliographic Records). These classes form building blocks for developing either application or domain-specific ontologies, with event-aware views for modeling different manifestations of a relationship (Hunter 2001). This tool provides graphical user interfaces and is free to download, but it is still an experimental prototype (Leuf 2006, 217-8), without support, and assumes users understand Java, RDF, and basic ontology and metadata principles.

WebOnto is a freely available Java applet coupled with a customized web server (LispWeb), which provides browsing, visualization and editing of knowledge models via the web. WebOnto is currently being used with ScholOnto (discussed above) and PlanetOnto, for search, retrieval, news feeds, alerts, and presentations of laboratory-related information.

The Kraft project outlines steps to building shared ontologies: ontology scoping, domain analysis, ontology formulation, and top-level ontology (Jones et al. 1998). However, their methodology lacks comprehensive evaluation of ontologies and is not applicable to global domains (Stucken-schmidt 2005, 68).

The Protégé opensource Ontology Editor provides two main ways of modeling ontologies, and
can export in various formats including OWL. Used extensively in clinical medicine and the biomedical sciences, Protégé covers the full range of development processes (Leuf 2006, 209-210).

**KAON** (KArlsruhe ONtology) offers a stable opensource, comprehensive tool suite for ontology creation, management, and a framework for building applications; it was designed for business applications requiring scalability and efficient reasoning capabilities (Leuf 2006, 213).

**Chimæra** is a system for creating and maintaining distributed web ontologies, as well as for merging ontologies and providing multidimensional diagnoses to identify problems (Leuf 2006, 210-211). Chimæra can load and export files in OWL, and is available opensource.

There are many possible variations in the ability of software to combine and relate ontologies; Klein provides a comparison of several different approaches (see table 1): SKC (Scalable Knowledge Composition), Chimæra, PROMPT, SHOE (Simple HTML Ontology Extensions), OntoMorph, metamodel, OKBC (Open Knowledge Base Connectivity) and layering. Of these, OntoMorph addresses the majority of the stated problems in combining ontologies. However, Klein states that “mismatches in expressiveness between languages is not solvable” and more comprehensive schemes need to be developed for interoperability of ontologies (Klein 2001).

### 3.4 Encoding

For computer applications to be able to use ontologies, they must be encoded in machine-readable languages: in particular, all implicit relations between concepts must be explicitly encoded. To enable interoperability between ontologies and query engines, we need to agree on standards for these encodings. As in any other area, there is some disagreement on what is the most useful path. OntologyWorks used the draft ISO (International Organization for Standardization) standard, SCL (Simple Common Logic), which has been superceded by the Common Logic Standard, currently under development (ISO 2006). Since OntologyWorks does not seek interoperability with the broader public (it is a commercial effort), their focus was on what was most efficient and effective for their needs. However, if this standard is adopted by the ISO, it will likely compete with OWL for wider ontology development. CyCorp developed its own language, CycL, for their powerful Cyc system; however, their opensource components (OpenCyc and ResearchCyc) provide translators to certain other languages, and the ability to export selectively in OWL (CyCorp 2002). Schematron, “a language for making assertions about patterns found in xml documents,” is based on the tree pattern uncovered in the marked-up document. It allows you to determine which variant of a language you are working with, as well as to verify that it conforms to a particular schema (Leuf 2006, 218). Schematron was published as a draft ISO standard in 2004.

<table>
<thead>
<tr>
<th>Language level mismatches</th>
<th>Issues</th>
<th>SKC</th>
<th>Chim.</th>
<th>PROMPT</th>
<th>SHOE</th>
<th>OntoM.</th>
<th>Metamodel</th>
<th>OKBC</th>
<th>Layering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntax</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Representation</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semantics</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expressivity</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ontology level mismatches</td>
<td>Paradigm</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concept description</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coverage of model</td>
<td>Scope of concepts</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Synonyms</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Homonyms</td>
<td>M</td>
<td>U</td>
<td>U</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Encoding</td>
<td>M</td>
<td>M</td>
<td>U</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Practical problems</td>
<td>Finding alignments</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Diagnosis of results</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Repeatability</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ontology versioning</td>
<td>Identification</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Change tracking</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Table of problems and approaches for combined use of ontologies

Legend  
A: Solves problem automatically  
U: Solutions suggested to user  
M: Provides mechanism for specifying solution (Klein 2001)
A proposed RDF Thesaurus Specification provides “conce...
schema is mapped to local schemas (Pazienza et al., 2004). The Bremen University Semantic Translator for Enhanced Retrieval (BUSTER) is a middleware of this same type, designed to access and integrate multiple ontologies which are based on a common vocabulary. The general top-level ontology it uses is based on simple Dublin Core with some added refinements. (Visser and Schuster 2002). Thus the user must commit to the basic generalized vocabulary that is used to define concepts in all the source ontologies, and is not presented with a specific domain view (Stuckenschmidt 2005, 199-207).

In contrast, the OBSERVER (Ontology Based System Enhanced With Relationships for Vocabulary hEterogeneity Resolution) system requires the user to select his terms from one of the ontologies it supports; the source material that ontology covers is then queried (Figure 4). If the results are not satisfactory, the user query is rewritten into the ontologies of other information sources in order to query other holdings (Mena et al. 2000).

OBSERVER uses synonyms, hypernyms, hyponyms, overlap, disjointedness and coverage to map between ontologies, storing these relations in a central repository to use for translating queries Stuckenschmidt, 2005, 192-198). In this manner, heterogeneous databases and ontologies are managed without the need for a single global ontology (Mena et al. 2000). MAFRA (the Ontology MApping FRamework) also is based on distributed mediation systems rather than a centralized one (Pazienza et al. 2004).

3.6 Costs

While ontologies offer benefits in terms of interoperability, browsing and searching, reuse, and structuring knowledge in a domain, the costs must be considered. Costs include construction, learning, cross-mapping, and maintenance and continual development of both the ontologies and the software (Menzies 1997). Information about cost is difficult to obtain, as most efforts are prototypes or commercial developments. Tim Berners-Lee, a major proponent of the Semantic Web, downplays the total cost, and fails to consider methodologies, depth of ontology, or even level of usability in his online assessment (Berners-Lee 2005). In a later article with others, however, this stance is modified somewhat by implying that
general web applications may only need lightweight ontologies; and recognition that in certain commercial applications, the use of powerful heavyweight ontologies will easily recoup the cost (Shadbolt et al. 2006). Recently a cost estimation approach has been developed (ONTOCOM; a detailed description is available in (Bontas and Mochol 2006) and an example of its application to a particular ontology (DILIGENT) is described (Bontas and Tempich 2005), though the actual results of the many formulas upon the various cost drivers are not included in this publication. These cost drivers include:

- Personnel factors: ontologist/domain expert capability & experience, language and tool experience, and personnel continuity
- Project factors: tool support, multi-site development, and required development schedule
- Reuse/maintenance factors: ontology understandability, domain/expert unfamiliarity, and complexity of evaluation, modifications, and translations

Development of an ontology requires a shared conceptualization by domain experts, users and designers (de Bruijn 2003, 5); this is not only difficult, but requires such a high initial investment, it will only be supportable where there is commercial interest (Stuckenschmidt and van Harmelen 2005, 249). While the initial cost of ontology implementation is frightening, one IBM researcher predicts the long term maintenance of an ontology to be 80% of the cost (Welty 2005). In a recent survey of 34 ontology engineering projects, half of which were commercial, all participants emphasized the resource-intensive nature of domain analysis and the lack of low barrier methods and tools (Simperl and Tempich 2006). The implications are that there must be a clear and pressing need for the benefits of ontological indexing and retrieval, sufficient to provide extensive funding or the dedicated volunteer labor of known and trusted professionals. From the limited survey of the landscape performed for this report, it appears that funding is currently available in medical fields, environmental research, national defense, and business applications. The educational field may contain sufficient volunteer experts, university support, and grant-funded development to make ontology development feasible for instructional materials.

To be able to effectively apply an ontology, much less change it, one must learn it, another time-consuming task. Apart from domain knowledge, the person encoding the document must have a level of understanding approaching that of a skilled knowledge engineer (Marshall and Shipman 2003). To expect the average citizen to have or develop the necessary knowledge and skill to coherently apply a domain ontology to a document is infeasible (Marshall 2004). If the users will not apply the ontologies, then the application of metadata to resources must be performed by the institution or service. Hence the users only bear the cost if they pay for the service, either directly or indirectly; this implies that ontologies may indeed only be feasible, in the long term, for applications in commercial services.

The only other solution to this cost would be the automation of application of ontologies to resources. The development of this functionality depends heavily on research and tools developed by the artificial intelligence community. Some of the techniques developed include a noun phrasing technique for concept extraction and concept association based on context, frequency and co-occurrence of terms (Chen 1999). However, precise meanings for every relation are necessary for automatic classification (Weinstein and Birmingham 1998). A 2003 assessment stated that there are a number of issues to be resolved before natural language can be understood by computers; and the majority of information present on the web is in natural language (Fensel 2003a). However, for technical fields with more structured terminology, a text-mining system for scientific literature, Textpresso, shows considerable promise for assisting in automatic ontology annotation. While the machine cannot replace the human expert, it can increase efficiency greatly (Müller et al. 2004). Further investigation into current developments in this area is warranted.

For the ontology to be widely usable and interoperable, cross-mapping to other ontologies and domains is necessary, requiring the involvement of multiple domain experts (Adam, Atluri, and Adiwijaya 2000). And ontologies (and their supporting software) must be expected to change (de Bruijn 2003, 35), as knowledge and terminology are continually evolving. It is quite possible that this aspect may restrict the usability of ontologies to specified domains. Cross-mapping is only likely if there is sufficient need and funding to offset the expense, and then it is not likely to be maintained over time without continued funding and demand.
4. Conclusions

The decision about if, when, and how one should apply ontologies to one’s digital library is a complex one. There are many aspects to consider, and several of those aspects are moving targets. Any assessment or survey, such as this one, can only be a snapshot of an evolving landscape, and as such, is useful primarily in helping one get his bearings for the moment. Further research and feasibility studies are necessary components for any digital library considering the application of ontologies.

Some purposes of ontologies may be particularly useful. Dieter Fensel predicted in 2003 that three areas in which ontology application potentially has a huge impact are knowledge management, enterprise integration, and e-commerce (Fensel 2003b). Already this prediction seems to be proving true. If one’s digital library falls into these domains, the usefulness may outweigh the cost: funding created by demand for a service may well be sufficient to overcome other obstacles. Usefulness in educational realms seems quite promising, but the return on investment has yet to be proven (Milam 2005).

Outside of heavily funded domains, feasibility is yet to be determined. If the target audience for the digital library is the general public, at no cost to the user, then it is not likely that the application of ontologies is currently monetarily feasible. Ontologies incur tremendous expenditures of resources in their creation or adoption, application, cross-mapping, maintenance, and possibly software development. Tools exist to assist in modifying existing ontologies, but they are not simple, and require extensive domain knowledge and understanding of the concepts and relations required for the ontology to be functional. Tools to apply ontologies to existing resources are still under development. Cross-mapping ontologies for use beyond a single domain is a new territory; if the source ontologies have the same basis, query engines appear to have good results, but that’s a rather telling caveat. Otherwise, it seems that only general mappings are feasible, supporting general queries with limited precision. To some extent, mappings can be automated, but must still be reviewed by a human.

Systems to support ontology use (query engines and semantic web browsers) are becoming available, but their usefulness is limited by the ontologies and their mappings. And the cost of maintenance and continual evolution of an ontology is yet unmeasured. On the other hand, a general ontology language has been adopted by W3C, tools and systems continue to evolve, and new ontologies appear every year. If funding exists, and an acceptable ontology exists in OWL for a domain covered by a particular digital library, it would be reasonable to assess the existing tools for application and delivery, and possibly move forward in implementation. As ontology and tool development lowers the technical and cost barriers, general digital libraries should certainly become involved: this is perhaps in the very near future.

References

Binding, Ceri and Tudhope, Douglas. 2004. KOS at your Service: programmatic access to knowledge


Klein, Michael. 2001. Combining and relating ontologies: an analysis of problems and solutions. In International Joint Conferences on Artificial Intelligence Workshop on Ontologies and Information


MIT and Hewlett Packard. 2007b DSpace: Welcome to DSpace. A joint project of Massachusetts Insti-


Automated Classification of Textual Documents Based on a Controlled Vocabulary in Engineering†

Koraljka Golub,* Thierry Hamon,** and Anders Ardö***

*KnowLib Research Group, Lund University, P. O. Box 118, SE-221 00 Lund, Sweden <Koraljka.Golub@eit.lth.se>

** Laboratoire d’Informatique de Paris-Nord – UMR CNRS 7030, Institut Galilée, Université Paris-Nord, Avenue J.-B. Clément, 93430 Villetaneuse, France <thierry.hamon@lipn.univ-paris13.fr>

***KnowLib Research Group, Lund University, P. O. Box 118, SE-221 00 Lund, Sweden <Anders.Ardo@eit.lth.se>

Koraljka Golub has interest in traditional and recent knowledge organization systems in the context of digital libraries. She acquired her doctorate from Lund University, Sweden in 2007. Her thesis dealt with automated subject classification in Web-based hierarchical browsing systems. From 2008 she will work as a research officer at UKOLN, Bath. One project will be about terminology registries, and the other on social tagging and ways it can enhance information retrieval, especially when combined with more traditional controlled vocabularies.

Thierry Hamon is assistant professor at the Computer Science Department, Paris-Nord University. He received his Ph.D. in computer science in 2000, on the topic of semantic variation in specialized corpora. His current research interest is in terminology acquisition and structuring, and bringing together tools for Natural Language Processing (NLP). He has developed several NLP tools: a terminological system SynoTerm dedicated to the acquisition of synonymy relations between terms, based on lexical resources, a term extractor, and a linguistic platform for the enrichment of specialized web documents.

Anders Ardö is Associate Professor at the Department of Electrical and Information Technology, Lund University, where managers the Knowledge Discovery and Digital Library Research Group (KnowLib). He has a background in Computer Systems with a PhD from Lund University in 1986. Since 1992 he has worked with research and development for digital library services. He has participated in many EU-projects including DESIRE, Telematics Applications Programme, Renardus, ALVIS and DELOS.

† Many thanks to Traugott Koch, Douglas Tudhope, Marianne Lykke Nielsen, and anonymous reviewers for providing comments on the manuscript, which helped improve the paper considerably. The authors also wish to thank two subject experts who helped in evaluation. This work was supported by the IST Programme of the European Community under ALVIS (IST-002068-STP).

Subject classification is organization of objects into topically related groups and establishing relationships between them. In automated subject classification (in further text: automated classification) human intellectual processes are replaced by, for example, statistical and computational linguistics techniques. Automated classification of textual documents has been a challenging research issue for several decades. Its relevance is rapidly growing with the advancement of the World Wide Web. Due to high costs of human-based subject classification and the ever-increasing number of documents, there is a danger that recognized objectives of bibliographic systems (Svenonius 2000, 20-21) would be left behind; automated means could provide a solution to preserve them (30).

Automated classification of text has many different applications (see Sebastiani 2002 and Jain et al. 1999); in this paper, the application context is that of information retrieval. In information retrieval systems, e.g., library catalogues or indexing and abstracting services, improved precision and recall are achieved by controlled vocabularies, such as classification schemes and thesauri. The specific aim of the classification algorithm is to provide a hierarchical browsing interface to a document collection, through a classification scheme. In our opinion, one can distinguish between three major approaches to automated classification: text categorization, document clustering, and document classification (Golub 2006a).

In document clustering, both subject clusters or classes into which documents are classified and, to a limited degree, relationships between them are automatically produced. Labeling the clusters is a major research problem, with relationships between them, such as those of equivalence, related-term and hierarchical relationships, being even more difficult to automatically derive (Svenonius 2000, 168). In addition, “[a]utomatically-derived structures often result in heterogeneous criteria for category membership and can be difficult to understand” (Chen and Du-
degrading recall. Medelyan and Witten (2006) showed how information from a subject-specific thesaurus improved performance of keyphrase extraction by more than 1.5 times in F1, precision, and recall.

The overall purpose of this experiment is to gain insights into what degree a good controlled vocabulary such as Engineering Information thesaurus and classification scheme (Milstead 1995) (in further text: Ei controlled vocabulary) could be used in automated classification of text, using string-matching. Vocabulary control in thesauri is achieved in several ways (Aitchinson et al. 2000). We believe that the following could be beneficial in the process of automated classification:

– Terms in thesauri are usually noun phrases, which are content words;
– Three main types of relationships are displayed in a thesaurus:
  - equivalence (e.g., synonyms, lexical variants);
  - hierarchical (e.g., generic, whole-part, instance relationships); and,
  - associative (terms that are closely related conceptually but not hierarchically and are not members of an equivalence set).
– In automated classification, equivalence terms could allow for discovering concepts and not just terms expressing the concepts. Hierarchies could provide additional context for determining the correct meaning of a term; and so could associative relationships;
– When a term has more than one meaning in the thesaurus, each meaning is indicated by the addition of scope notes and definitions, providing additional context for automated classification.

In a previous paper Golub (2006c) explored to what degree different types of Ei thesaurus terms and Ei classification captions influence performance of automated classification. In short, the algorithm searched for terms from the Ei controlled vocabulary in engineering documents to be classified (see 2.1). The majority of classes were found when using all the types of terms: preferred terms, their synonyms, related, broader, narrower terms and captions, in combination with a stemmer: recall was 73%. The remaining 27% of classes were not found because the words in the term list designating the classes did not exist in the text of the documents to be classified. No weighting or cut-offs were applied in the experiment. Apart from showing that all those types of terms should be used for a term list in order to achieve best recall, it was also indicated that higher weights could be given to preferred terms (from the thesaurus), captions (from the classification scheme) and synonyms (from the thesaurus), as those three types of terms yielded highest precision.

The aim of this experiment is to improve the classification algorithm based on string-matching between the Ei controlled vocabulary and engineering documents to be classified. We especially wanted to do the following:

– increase levels of F1 and precision, similar to those of recall from the previous experiment (Golub 2006c, 964), by applying different weights and cut-offs; and,
– increase levels of recall to more than those achieved in the previous experiment by adding new terms extracted using natural language processing methods such as multi-word morpho-syntactic analysis and synonym extraction.

2. Methodology

2.1 String matching algorithm

This section describes the classification algorithm used in the experiment. It is based on searching for terms from the Ei controlled vocabulary, in the field of engineering, in text of documents to be classified (also in the field of engineering). The Ei controlled vocabulary consists of two parts: a thesaurus of engineering terms, and a hierarchical classification scheme of engineering topics. These two controlled vocabulary types have each traditionally had distinct functions: the thesaurus has been used to describe a document with as many controlled terms as possible, while the classification scheme has been used to group similar documents together to the purpose of shelving them and allowing systematic browsing. The aim of the algorithm was to classify documents into classes of the Ei classification scheme in order to provide a browsing interface to the document collection. A major advantage of Ei is that thesaurus descriptors are mapped to classes of the classification scheme. These mappings have been made manually (intellectually) and are an integral part of the thesaurus. Compared with captions alone, mapped thesaurus terms provide a rich additional vocabulary for every class: instead of having only one term per class (there is only one caption per class), in our experiment there were on average 88 terms per class. (A caption is a class notation expressed in words, e.g., in
the Ei classification scheme “Electric and Electronic Instruments” is the caption for class “942.1.”

Pre-processing steps of Ei included normalizing upper- and lower-case words. Upper-case words were left in upper case in the term list, assuming that they were acronyms; all other words containing at least one lower-case letter were converted into lower case. The first major step in designing the algorithm was to extract terms from Ei into what we call a term list. It contained class captions, thesaurus terms (Term), classes to which the terms and captions map or denote (Class), and weight indicating how appropriate the term is for the class to which it maps or which it designates (Weight). Geographical names, all mapping to class 95, were excluded on the grounds that they are not engineering-specific. The term list was formed as an array of triplets:

**Weight**: Term (single word, Boolean term or phrase) = Class

*Single-word terms* were terms consisting of one word. *Boolean terms* were terms consisting of two or more words that must all be present but in any order or in any distance from each other. Boolean terms in this form were not explicitly part of Ei, but were created to our purpose. They were considered to be those terms which in Ei contained the following strings: *and*, *vs.* (short for *versus*), , (comma), ; (semi-colon, separating different concepts in class captions), ( and ) (parentheses, indicating the context of a homonym), : (colon, indicating a more specific description of the previous term in a class captions), and – (double dash, indicating *heading-subheading* relationship). These strings we replaced with @and which indicated the Boolean relation in the term. All other terms consisting of two or more words were treated as *phrases*, i.e., strings that need to be present in the document in the exact same order and form as in the term. Ei comprises a large portion of composite terms (3,474 in the total of 4,411 distinct terms in our experiment); as such, Ei provides a rich and precise vocabulary with the potential to reduce the risks of false hits.

The following are two excerpts from the Ei classification scheme and thesaurus, based on which the excerpt from the term list (further below) is created:

**From the classification scheme:**

931.2 Physical Properties of Gases, Liquids and Solids

...  
942.1 Electric and Electronic Instruments

...  
943.2 Mechanical Variables Measurements

**From the thesaurus:**

TM Amperometric sensors

UF Sensors–Amperometric measurements

MC 942.1

...

TM Angle measurement

UF Angular measurement

UF Mechanical variables measurement–Angles

BT Spatial variables measurement

RT Micrometers

MC 943.2

...

TM Anisotropy

NT Magnetic anisotropy

MC 931.2

All the different thesaurus terms as well as captions were added to the term list. Despite the fact that choosing all types of thesaurus terms might lead to precision losses, we decided to do just that in order to achieve maximum recall, as shown in a previous paper (Golub 2006c). In the thesaurus, TM stands for the preferred term, UF (“Used For”) for an equivalent term, BT for broader term, RT for related term, NT for narrower term; MC represents the main class; sometimes there is also OC, which stands for optional class, valid only in certain cases. Main and optional classes are classes from the Ei classification scheme that have been made manually (intellectually) and are an integral part of the thesaurus. Based on the above excerpts, the following term list would be created:

1: physical properties of gases @and liquids @and solids = 931.2,
1: electric @and electronic instruments = 942.1,
1: mechanical variables measurements = 943.2,
1: amperometric sensors = 942.1,
1: sensors @and amperometric measurements = 942.1,
1: angle measurement = 943.2,
1: angular measurement = 943.2,
1: mechanical variables measurement @and angles = 943.2,
1: spatial variables measurement = 943.2,
1: micrometers = 943.2,
1: anisotropy = 931.2,
1: magnetic anisotropy = 931.2,
The number at the beginning of each triplet is weight estimating the probability that the term of the triplet designates the class; in this example it is set to 1 as a baseline, and experiments with different weights are discussed later on.

The algorithm searches for strings from a given term list in the document to be classified and if the string (e.g., magnetic anisotropy from the above list) is found, the class(es) assigned to that string in the term list (931.2 in our example) are assigned to the document. One class can be designated by many terms, and each time a term is found, the corresponding weight (1 in our example) is added to a score for the class. The scores for each class are summed up and classes with scores above a certain cut-off (heuristically defined, discussed later on) are selected as the final ones for the document being classified.

The Ei classification scheme is hierarchical and consists of six main classes divided into 38 finer classes which are further subdivided into 182 classes. These are subdivided even further, resulting in some 800 individual classes in a five-level hierarchy. For this experiment one of the six main classes was selected, together with all its subclasses: class 9, Engineering, General. The reason for choosing this class was that it covers both natural sciences such as physics and mathematics, and social sciences fields such as engineering profession and management. The literature of the latter tends to contain more polysemic words than the former, and as such presents a more complex challenge for automated classification.

Within the 9 class, there are 99 subclasses. However, for seven of them the number of documents in a database based on which the document collection was created (see 2.2 Document collection) were few, less than 100. Thus those seven classes were excluded from the experiment altogether. These were: 9 (Engineering, General), 902 (Engineering Graphics; Engineering Standards; Patents), 91 (Engineering Management), 914 (Safety Engineering), 92 (Engineering Mathematics), 93 (Engineering Physics), and 94 (Instruments and Measurement). Of the remaining 92 classes, the distribution at the five different hierarchical levels is as follows: at the fifth hierarchical level 11 classes, at the fourth 67, at the third 14, and at the second hierarchical level 5.

2.2 Document collection

The document collection comprised 35,166 bibliographic records from the Compendex database (Engineering Information 2006). (Compendex being a commercial database, the document collection cannot be made available to others, but the authors are willing to provide documents’ identification numbers on request.) The records were selected by simply retrieving the top 100 or more of them upon entering the class notation. A minimum of 100 records per class were downloaded at several different points in time during the years of 2005 and 2006.

For each record there was at least one of the 92 selected classes that were human-assigned (see 2.1). A subset of this collection was created to include only those records where main class was class 9 (The first one listed in the Ei classification codes field of the record.); this subset contained 19237 documents.

From each bibliographic record (in further text: document) the following elements were extracted: an identification number, title, abstract and human-assigned classes (Ei classification codes). Thesaurus descriptors (in Compendex called Ei controlled terms) were not extracted since the purpose of this experiment was to compare automatically assigned classes (and not descriptors) against the human-assigned ones. Below is an example of one document:

Identification number: 03337590709
Title: The concept of relevance in IR
Abstract: This article introduces the concept of relevance as viewed and applied in the context of IR evaluation, by presenting an overview of the multidimensional and dynamic nature of the concept. The literature on relevance reveals how the relevance concept, especially in regard to the multidimensionality of relevance, is many faceted, and does not just refer to the various relevance criteria users may apply in the process of judging relevance of retrieved information objects. From our point of view, the multidimensionality of relevance explains why some will argue that no consensus has been reached on the relevance concept. Thus, the objective of this article is to present an overview of the many different views and ways by which the concept of relevance is used - leading to a consistent and compatible understanding of the concept. In addition, special attention is paid to the type of situational relevance. Many researchers perceive situational relevance as the most realistic type of user relevance, and therefore situational relevance is discussed with reference to its potential dynamic nature, and as a requirement for interactive information retrieval (IIR) evaluation.
**Ei classification codes:** 903.3 Information Retrieval & Use, 723.5 Computer Applications, 921 Applied Mathematics

Automated classification was based on title and abstract, and automatically assigned classes were compared against human-assigned ones (Ei classification codes in the example). On average, 2.2 classes per document were human-assigned, ranging from 10 to 1.

### 2.3 Evaluation methodology

#### 2.3.1 Evaluation challenge

According to ISO standard on methods for examining documents, determining their subjects, and selecting index terms (International Standards Organization 1985), human-based subject indexing is a process involving three steps: 1) determining subject content of a document, 2) conceptual analysis to decide which aspects of the content should be represented, and 3) translation of those concepts or aspects into a controlled vocabulary. These steps, in particular the second one, are based on a specific library’s policy in respect to its document collections and user groups. Thus, when evaluating automatically assigned classes against the human-assigned ones, it is important to know the human-based indexing policies. Unfortunately, we were unable to obtain indexing policies applied in the Compendex database. What we could derive from the document collection was the number of human-assigned classes per document, which were used in evaluation. However, without a thorough qualitative analysis of automatically assigned classes one cannot be sure whether, for example, the classes assigned by the algorithm, but not human-assigned, are actually wrong, or if they were left out by mistake or because of the indexing policy. A further issue is that we did not know whether the articles had been human-classified based on their full-text or/and abstracts; we had, however, only abstracts.

Another problem to consider when evaluating automated classification is the fact that certain subjects are erroneously assigned. When indexing, people make errors such as those related to exhaustivity policy (too many or too few terms become assigned), specificity of indexing (which usually means that people do not assign the most specific term), they may omit important terms, or assign an obviously incorrect term (Lancaster 2003, 86-87). In addition, it has been reported that different people, whether users or professional subject indexers, would assign different subject terms or classes to the same document. Studies on inter- and intra-indexer consistency report generally low indexer consistency (Olson and Boll 2001, 99-101). Markey (1984) reviewed 57 indexer consistency studies and reported that consistency levels range from 4% to 84%, with only 18 studies showing over 50% consistency. There are two main factors that seem to affect it:

1. Higher exhaustivity and specificity of subject indexing both lead to lower consistency, i.e., indexers choose the same first term for the major subject of the document, but the consistency decreases as they choose more classes or terms;
2. The bigger the vocabulary, or, the more choices the indexers have, the less likely will they choose the same classes or terms (Olson and Boll 2001, 99-101).

Both of these two factors were present in our experiment:

1. High exhaustivity: on average, 2.2 classes per document had been human-assigned, ranging from 10 to 1;
2. Ei controlled vocabulary is rather big (we chose 92 classes) and deep (five hierarchical levels), allowing many different choices.

An analysis of automatically and human-assigned classes in a previous study showed, among other things, how certain human-assigned classes were actually wrong and some automatically-assigned classes that were not human-assigned were correct (Golub 2006b). An analysis conducted within this study proved the same (see section 4.3).

Today evaluation in automated classification experiments is mostly conducted under controlled conditions, ignoring the above-discussed issues. As Sebastiani (2002, 32) puts it:

> The evaluation of document classifiers is typically conducted experimentally, rather than analytically. The reason is that ... we would need a formal specification of the problem that the system is trying to solve (e.g., with respect to what correctness and completeness are defined), and the central notion ... that of membership of a document in a category is, due to its subjective character, inherently nonformalizable.
Because of the fact that methodology for such experiments has yet to be developed, as well as limited resources, we followed the common approach to evaluation and started from the assumption that human-assigned classes in the document collection were correct, and compared automatically assigned classes against them.

2.3.2 Evaluation measures

The subset of the Ei controlled vocabulary we used comprised 92 classes that are all topically related to each other. The topical relatedness is expressed in numbers representing the classes: the more initial digits any two classes have in common, the more related they are. For example, 933.1.2 for *Crystal Growth* is closely related to 933.1 for *Crystalline Solids*, both of which belong to 933 for *Solid State Physics*, and finally to 93 for *Engineering Physics*. Each digit represents one hierarchical level: class 933.1.2 is at the fifth hierarchical level, 933.1 at the fourth etc. Thus, comparing two classes at only first few digits (later referred to as partial matching) instead of all the five also makes sense. Still, unless specifically noted, the evaluation in this experiment was conducted based on all the five different levels (later referred to as complete matching) instead of all the five also makes sense. Still, unless specifically noted, the evaluation in this experiment was conducted based on all the five different levels (later referred to as complete matching), i.e., an automatically assigned class was considered correct only if all its digits were the same as a human-assigned class for the same document.

Evaluation measures used were the standard microaveraged and macroaveraged precision, recall and F1 (Sebastiani 2002, 40-41), for both complete and partial matching:

\[
\text{Precision} = \frac{\text{correctly automatically assigned classes}}{\text{all automatically assigned classes}}
\]

\[
\text{Recall} = \frac{\text{correctly automatically assigned classes}}{\text{all human-assigned classes}}
\]

\[
F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

In macroaveraging the results are first calculated for each class, and then summed and divided by the number of classes. In microaveraging the results for each part of every equation are summed up first (e.g., all correctly automatically assigned classes are added together, all automatically assigned classes are added together), and then the “aggregated” values are used in one equation. Equations for macroaveraged and microaveraged precision are given below:

\[
\text{Precision}_{\text{macroaveraged}} = \frac{\text{sum of precision values for each class}}{\text{number of all classes}}
\]

\[
\text{Precision}_{\text{microaveraged}} = \frac{\text{sum of correct automated assignments for each class}}{\text{sum of all automated assignments for each class}}
\]

In microaveraging more value is given to classes that have a lot of instances of automatically assigned classes and the majority of them are correct, while in macroaveraging the same weight is given to each class, no matter if there are many or few automatically assigned instances of it. The differences between macroaveraged and microaveraged values can be large, but whether one is better than the other has not been agreed upon (Sebastiani 2002, 41-42). Thus, in this experiment, it is the mean macroaveraged and microaveraged F1 that is mostly used.

In order to examine different aspects of the automated classification performance, several other factors were also taken into consideration:

- Whether the (human-assigned) main class is found;
- The number of documents that got automatically assigned at least one class;
- Whether the class with highest score was the same as the human-assigned main class;
- The distribution of automatically versus human-assigned classes; and,
- The average number of classes assigned to each document. There were 2.2 human-assigned classes per document, and our aim was to achieve similar.

In the context of hierarchical browsing based on a classification scheme, having too many classes assigned to a document would place one document to too many different places, which would create the opposite effect of the original purpose of a classification scheme, that of grouping similar documents together.

3. Improving the algorithm

The major aim of the experiment was to improve the algorithm that was previously experimented with in Golub 2006c, where highest (microaveraged) recall was 73% when all types of terms were included in the term list. In that experiment neither weights nor cut-offs were experimented with, so all the classes that were found for a document were assigned to it. Here we wanted to achieve as high as possible precision levels by use of term weighting and class cut-offs. In or-
der to also allow for better recall, the basic term list was enriched with new terms extracted from documents in the Compendex database, using multi-word morpho-syntactic analysis and synonym acquisition.

3.1 Term weights

The aim of this part of the experiment was to achieve as high as possible precision levels by use of weighting and cut-offs. As shown in Golub 2006c, all types of terms need to be used in the term list for maximum recall. Thus, all the different types of terms and their mappings to classes were merged into the final term list. This resulted in a number of duplicate cases which were dealt with in the following manner:

- If one term mapping to the same class was a caption, a preferred term, and a synonym at the same time, the highest preference was, based on their performance (see Table 4), given to captions, followed by preferred terms, followed by synonyms, while others were removed from the list;
- If one term mapping to both optional class (OC) and main class (MC) was a caption, a preferred term, and a synonym at the same time, the highest preference was, based on their performance (see Table 4), given to captions, followed by preferred terms, followed by synonyms, while others were removed from the list;
- If one thesaurus term of the same type mapped to both optional class (OC) and main class (MC), the one that mapped to the optional class was removed (based on their performance, see Table 2).

The final term list consisted of 8099 terms, out of which 92 were captions (all mapped to main class (MC)), 668 were broader terms, 729 narrower, 3224 related, and 1733 were synonym terms. This big number of terms that have been human-mapped to classes indicates potential usefulness of such a controlled vocabulary in a string-matching algorithm for automated classification.

In order to systematically vary different parameters, the following 14 weighting schemes evolved:

1. w1: All terms in the term list were given the same weight, 1. This term list served as a baseline.

2. w134: Different term types were given different weights: single-word terms 1, phrases 3, and Boolean terms 4. These weights were heuristically derived in a separate experiment (Table 1). Three different term lists were created, each containing only single-word terms, phrases or Boolean terms. Weight 1 was assigned to all of them. The documents were classified using these three terms lists and their performance was compared for precision.

<table>
<thead>
<tr>
<th></th>
<th>Single</th>
<th>Phrase</th>
<th>Boolean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. precision (%)</td>
<td>8</td>
<td>26</td>
<td>33</td>
</tr>
<tr>
<td>Derived weight</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 1. Single, phrase and Boolean term lists and their performance as a basis for weights.

Avg. precision (%) is mean microaveraged and macroaveraged precision. Derived weights were based on dividing precision values (Avg. precision) by the lowest precision value (in this case 8).

3. w12: Terms mapping to a main class (MC) were given weight 2, and those mapping to an optional class (OC) were given weight 1. These weights were heuristically derived in a separate experiment (Table 2). Two different term lists were created, one containing only those terms that map to a main class, and another one containing only those terms that map to an optional class. Weight 1 was assigned to all of them. The documents were classified using these two terms lists and their performance was compared for precision.

<table>
<thead>
<tr>
<th></th>
<th>MC</th>
<th>OC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. precision (%)</td>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td>Derived weight</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2. Main code and optional code term lists and their performance as a basis for weights.

Avg. precision (%) is mean microaveraged and macroaveraged precision. Derived weights were based on dividing precision values (Avg. precision) by the lowest precision value (in this case 6).

4. w134_12: This list was a combination of the two preceding lists. Weights for term type 1, 3, and 4 for single, phrase or Boolean term were multiplied by the weight for the type of class to which the term mapped – 1 or 2 for optional or main class.

5. wOrig: As used in the original term weighting scheme when the string-matching algorithm based on EI was first applied (Koch and Ardö 2000). These weights were intuitively derived. They combined types of terms depending if it were a single-word term, Boolean or phrase, and whether the assigned class was main (MC) or optional (OC).
Table 3. Weights in the original algorithm.

<table>
<thead>
<tr>
<th>OC</th>
<th>Phrase</th>
<th>Boolean</th>
<th>Single</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

6. **w1234**: With weights for different Ei term type as experimented with in Golub 2006c (captions are from the classification scheme, all others are thesaurus terms).

7. **w134_1234**: This list was a combination of two previous lists, w134 and w1234. Weights for term type 1, 3, and 4 for single, phrase or Boolean term were multiplied by the weight for the type of Ei term as given in Table 4.

8. **w134_12_1234**: This list was a combination of two previous lists, w134_12 and w1234. Weights for term type 1, 3, and 4 for single, phrase or Boolean term were multiplied by the weight for the type of class to which the term mapped – 1 or 2 for optional or main class, and by the weight for the type of Ei term as given in Table 4.

9. **wTf10**: In this list weights were based on the number of words the term consisted of, and of the number of times each of its words occurred in other terms (cf. *tf-idf*, term frequency – inverse document frequency, Salton and McGill 1983, 63, 205). If \( f \) were the frequency with which a word \( w \) from the term \( t \) occurred in other terms, term \( t \) consisting of \( n \) words, then the weight of that term was calculated as follows:

\[
\text{weight}_t = \log(n) \cdot \left( 1/f_{w_1} + 1/f_{w_2} + \ldots + 1/f_{w_n} \right)
\]

Logarithm was applied in order to reduce the impact of parameter \( n \), i.e., to avoid getting overly high weights for terms consisting of several sparse words. In order to get integers as weights, the weights were multiplied by 10, rounded and increased by 1 to avoid zeros.

10. **wTf10Boolean**: As in wTf10, with all the phrases modified into Boolean terms. This list was created in order to study the influence of phrases and Boolean terms on precision and recall.

11. **wTf10Phrases**: As in wTf10, with all the Boolean terms modified into phrases. This list was created in order to study the influence of phrases and Boolean terms on precision and recall.

12. **wTf10_12**: As in wTf10, with those weights multiplied by the weight for the type of class to which the term maps – 1 or 2 for optional or main class. The multiplication was done before the rounding.

13. **wTf10_1234**: As in wTf10, with those weights multiplied by the weight for the type of relationship (Table 4). The multiplication was done before the rounding.

14. **wTf10_12_1234**: As in wTf10_12, with those weights multiplied by the weight for the type of relationship (Table 4). The multiplication was done before the rounding.

3.1.1 Stop-word list and stemming

Although the terms and captions in the Ei controlled vocabulary are usually noun phrases which are good content words, they can also contain words which are frequently used in many contexts and as such are not very indicative of any document’s topicality (e.g., word *general* in the Ei class caption *Engineer*, *General*). Thus, a stop-word list was used. It contained 429 such words, and was taken from Onix text retrieval toolkit (Onix text retrieval toolkit). For stemming, the Porter’s algorithm (Porter 1980) was used. The stop-word list was applied to the term lists, and stemming to the term lists as well as documents.

3.2 Cut-offs

In a previous experiment (Golub 2006c) cut-offs were not used—instead, all the classes that were found for a document were assigned to it. In the context of hierarchical browsing based on a classification scheme, having too many classes assigned to a document would place one document to many different places, which would create the opposite effect of the original purpose of a classification scheme (grouping similar documents together). In the
document collection, there were 2.2 human-assigned classes per document, and the aim of automated classification was to achieve similar. The effect of several different cut-offs was investigated:

1. All automatically derived classes are assigned as final ones (no cut-off).
2. In order to assign a certain class as final, the score of that class had to have a minimum percentage of the sum of all the classes' scores. Different values for the minimum percentage were tested: 1, 5, 10, 15 and 20, as well as some others (see section 4 Results).
3. The second type of cut-off in combination with the rule that if there were no class with the required score, the one with the highest score would be assigned.
4. In order to follow the subject classification principle of always assigning the most specific class possible, the principle of score propagation was introduced. The principle was implemented so that the scores for classes at deeper hierarchical levels were a sum of their own score together with scores of classes at upper hierarchical levels if such were assigned.

3.3 Enriching the term list with new terms

In the previous experiment (Golub 2006c), highest achieved recall was 73% (microaveraged), when all types of terms were included in the term list. In order to further improve recall, the basic term list was enriched with new terms. These terms were extracted from bibliographic records of the Compendex database, using multi-word morpho-syntactic analysis and synonym acquisition, based on the existing preferred and synonymous terms (as they gave best precision results).

Multi-word morpho-syntactic analysis was conducted using a parser FASTER (Jacquemin 1996) which analyses raw technical texts and, based on built-in meta-rules, detects morpho-syntactic variants. The parser exploits morphological (derivational and inflectional) information as given by the database CELEX (Baayen et al. 1995). Morphological analysis was used to identify derivational variants, such as:

- effect of gravity: gravitational effect
- architectural design: design of the proposed architecture
- supersonic flow: subsonic flow
- structural analysis: analysis of the structure

Syntactical analysis was used to:

- a) insert word inside a term, such as:
  - flow measurement: flow discharge measurements
  - distribution of good: distribution of the finished goods
  - construction equipment: construction related equipment
  - intelligent control: intelligent distributed control
- b) permute components of a term, such as:
  - control of the inventory: inventory control
  - flow control: control of flow
  - development of a flexible software: software development
- c) add a coordinated component to a term, such as:
  - project management: project schedule and management
  - control system: control and navigation system

Synonyms were acquired through a rule-based system SynoTerm (Hamon and Nazarenko 2001) which infers synonymy relations between complex terms by employing semantic information extracted from lexical resources. First the documents were preprocessed and tagged with part-of-speech information and lemmatized. Then terms were identified through the YiTeA term extractor (Aubin and Hamon 2006). The semantic information provided by the database WordNet (Fellbaum 1998; WordNet) was used as a bootstrap to acquire synonym terms of the basic terms. The synonymy of the complex candidate terms was assumed to be compositional, i.e., two terms were considered synonymous if their components were identical or synonymous (e.g., building components: construction components, building components: construction elements).

Although verification by a subject expert is desirable for all automatically derived terms, due to limited resources only the extracted synonyms were verified. Checking the synonyms is also most important since computing those leads to a bigger semantic shift than morphological and syntactical operations do. The verification was conducted by a subject expert, a fifth-year student of engineering physics. Suggested synonym terms were displayed in the user interface of SynoTerm. The verification was not strict: derived terms were kept if they were semantically related to the basic term. Thus, hyperonym
(generic/specific) or meronym (part/whole) terms were also accepted as synonyms. The expert spent 10 hours validating the derived terms. Of the 292 automatically acquired synonyms, 168 (57.5%) were validated and used in the experiment.

4. Results

4.1 Improving F1 and precision: applying weights and cut-offs

Based on each of the 14 term lists, the classification algorithm was run on the document collection of 35,166 documents (see 2.2). As described earlier (2.3.2), several aspects were evaluated and different evaluation measures were used; thus, for each term list, the following types of results were obtained:

1. min 1: if no classes were assigned because their final scores were below the pre-defined cut-off value (described in 3.2), the class with the highest score was assigned;
2. cut-off: the applied cut-off value;
3. min 1 correct: number of documents that were assigned at least one correct class;
4. min 1 auto: number of documents that were assigned at least one class;
5. avg auto/doc: average number of classes that were assigned per document, based on documents that were assigned at least one class;
6. macroa P: macroaveraged precision;
7. macroa R: macroaveraged recall;
8. macroa F1: macroaveraged F1;
9. microa P: microaveraged precision;
10. microa R: microaveraged recall;
11. microa F1: microaveraged F1;
12. mean F1s: arithmetic mean of macroaveraged and microaveraged F1 values.

The same experiment was run on all the 14 term lists. For each term list, two parameters were varied: 1) whether min 1 was assigned or not; and, 2) the first two cut-off variants from section 3.2.

When looking at mean F1 values, the differences between the term lists are not larger than four percent. Performance of the different lists measured in precision and recall is also similar. Three lists that perform best in terms of mean F1 are w1234, w134_1234 and w134_12_1234. Every list was run against stop-words removed, stemming, and both the stop-words removed and stemming, each in combination with different cut-off values: 5, 10 and 15. Improvements when using either stemming or stop-words removal or both are achieved in majority of cases up to two percent. There is also a slight increase in the number of correctly found classes without finding more incorrect classes. The differences between the three term lists measured in mean F1 are minor – one or two percent. The best term list is w134_12_1234 used in combination with stemming and stop-words removal and cut-off 10 – best mean F1 is 0.24. For this list more cut-offs were experimented with for better results; the value of 9 proved to perform best but better only on a third decimal digit than that of 10. In the following experiments, unless specifically noted, we used the best-performing w134_12_1234 term list and setting (applying stemming and stop-words removal, cut-off 9).

4.1.1 Stop words: removal and stemming

Next, the influence of stop-words removal and stemming was tested (as described in 3.1.1). For this experiment three lists that performed best in the previous one were chosen: w1234, w134_1234 and w134_12_1234. Every list was run against stop-words removed, stemming, and both the stop-words removed and stemming, each in combination with different cut-off values: 5, 10 and 15. Improvements when using either stemming or stop-words removal or both are achieved in majority of cases up to two percent. There is also a slight increase in the number of correctly found classes without finding more incorrect classes. The differences between the three term lists measured in mean F1 are minor – one or two percent. The best term list is w134_12_1234 used in combination with stemming and stop-words removal and cut-off 10 – best mean F1 is 0.24. For this list more cut-offs were experimented with for better results; the value of 9 proved to perform best but better only on a third decimal digit than that of 10. In the following experiments, unless specifically noted, we used the best-performing w134_12_1234 term list and setting (applying stemming and stop-words removal, cut-off 9).

4.1.2 Individual classes

It was shown that certain classes perform much better than the average. Performance of different classes varies quite a lot. For example, top three performing classes as measured in precision are different from top three classes for recall or F1: see Table 5.

4.1.3 Partial matching

As expected, the algorithm performs better when evaluation is based on partial matching between automatically and human-assigned classes (see sec-
As seen from Table 6, at the second hierarchical level F1 is up to 0.66 and at third 0.59. At the second hierarchical level the best F1 is achieved by classes Engineering mathematics (represented by notation 92) and General engineering (90). At the third hierarchical level, the class that performs best of all is 921 Applied Mathematics, while the worst one is 943 Mechanical and Miscellaneous Instruments. In conclusion, for the 14 classes at top three hierarchical levels mean F1 is almost twice as good as for the complete matching, which implies that our classification approach would suit better those information systems in which fewer hierarchical levels are needed, like the Intute subject gateway on engineering (Intute Consortium 2006).

The variations in performance between individual classes for both complete and partial matching are quite big, but at this stage it is difficult to say why. The two best-performing classes at the second hierarchical level have by far the smallest number of terms designating them (terms). However, in other cases there does not seem to be any correlation between number of terms and performance, as also discovered in Golub 2006b. Further research is needed to explore what the factors contributing to performance are.

### 4.1.4 Score propagation

A relevant subject classification principle is to always assign the most specific class available. This principle provided us with a basis for score propagation, in which scores of classes at narrower (more specific) hierarchical levels were increased by scores assigned to their broader classes (later referred to as “propagated down”). In another run, this was slightly varied, so that the broader classes from which scores were propagated to their narrower classes were removed (“propagated down, broader removed”).

These types of score propagation were tested on the best performing term list and setting (w134_12-1234 with stemming and stop-words removal). In complete matching, “propagated down” performs best. However, it is slightly worse than when not using score propagation at all. In partial matching, both “propagated down” and “propagated down, broader removed” perform slightly better than the original on the first two or three hierarchical levels, and slightly worse on the fourth and fifth ones. These not-so-good results with score propagation can be partially explained by the fact that the term list contained both broader and narrower terms, which was done in order to achieve best recall (Golub 2006c).

### 4.1.5 Finding main classes

We further analyzed the degree to which the one most important concept of every document is found by the algorithm. To this purpose, a subset of (19,153) documents was used which had the human-assigned main class in class 9 (there is one main class per document). In complete matching, 78% of main classes are found when no cut-offs are applied. When cut-offs are applied, 22% of main classes are found. In partial matching, more main classes are found at the second and third hierarchical levels when using both types of score propagation, up to 59% and 38% respectively. Thus, score propagation could be used in services for which fewer hierarchical levels are needed (e.g., Intute Consortium 2006).
4.1.6 Distribution of classes

Using the same best setting achieved so far, the algorithm was also evaluated for distribution of automatically assigned classes in comparison to that of the human-assigned ones. The comparison was based on how often two classes get assigned together when using the algorithm in comparison to when they get human-assigned. Figure 1 shows the frequency distribution of assigned class pairs. The x-coordinate presents human-assigned class pairs ordered by descending frequency. One point represents one class pair: e.g., the pair of classes 912.2 and 903 occurs most frequently in human-based classification (48 times, as marked on the y-coordinate) and is represented by point 1 on the x-coordinate; point 500 on the x-coordinate represents the 913.5 and 911 pair that occurs 3 times, as marked on the y-coordinate. Thus, the smoothest line (Human-assigned) represents the human-assigned classes. The minimum of 2538 pairs of classes that both the algorithm and people have produced are shown.

A correlation of 0.38 exists between the human-assigned classes and automatically assigned classes (Automated). However, for the 100 most frequent pairs, the correlation drops to 0.21. In the top 10 most frequent pairs of classes, there is no overlap at all. In conclusion, the distribution of human-assigned and automatically assigned classes is more correlated when looking at all pairs of classes occurring together, but less so for more frequently occurring pairs.

4.1.7 Implications for application

Since automated classification algorithms can have a number of different applications, it is important to emphasize that an algorithm can be adjusted for the specific application need. Here those applications are pointed out in which our algorithm was shown to yield promising results in terms of F1 and precision.

1. In all applications, best precision and F1 are achieved when applying the w134_12_1234 term list, together with stemming and stop-words removal.
2. In information systems such as Intute (Intute Consortium 2006), several broader hierarchical levels are used. To the purpose of such an application, the classification algorithm should be implemented so that only classes from top three hierarchical levels are used, but so that scores from classes at lower hierarchical levels are added to the final some of their broader classes.
3. In applications where classes at all hierarchical levels are needed, such as other hierarchical browsing systems, searching or machine-aided indexing softwares, cut-off level of nine, and the principle of assigning at least the class with highest score should be implemented. In addition, the choice can be made to assign only the class with highest score,
4.2 Enhancing the term list with new terms

In the previous experiment (Golub 2006c), highest achieved recall was 73% (microaveraged), when all types of terms were included in the term list. In order to further improve recall, the basic term list was enriched with new terms. These terms were extracted from bibliographic records of the Compendex database, using multi-word morpho-syntactic analysis and synonym acquisition, based on the existing preferred and synonymous terms (as they gave best precision results). The number of terms added to the term list was as follows:

1. Based on multi-word morpho-syntactic analysis:
   – derivation: 705, out of which 93 adjective to noun, 78 noun to adjective, and 534 noun to verb;
   – permutation: 1373;
   – coordination: 483;
   – insertion: 742; and
   – preposition change: 69.
2. Based on semantic variation (synonymy): 292 automatically extracted, out of which 168 were verified as correct by the subject expert.

In order to examine the influence of different types of extracted terms, nine different term lists were created and the classification was based on each of them. It was shown that the number of terms is not proportional to performance, e.g., permutation-based extraction comprises 1373 terms, and, when stemming is applied, has performance as measured in mean F1 of 0.02, whereas coordination comprises 403 terms, with performance of 0.07. These two cases can be explained by the fact that permutation also implies variation based on insertion and preposition change (e.g., engineering for commercial window systems: system engineering) which leads to bigger semantic shift than the identification of term variant based on the coordination. By combining all the extracted terms into one term list, the mean F1 is 0.14 when stemming is applied, and microaveraged recall is 0.11, which would imply that enriching the original Ei-based term list with these newly extracted terms should improve recall. In comparison to results gained in Golub 2006c, where microaveraged recall with stemming is 0.73, here the best recall, also microaveraged and with stemming, is 0.76.

The next step was to assign appropriate weights to the newly extracted terms (Table 7). We used the w14_12_1234 term list, earlier shown to perform best. The result as measured in mean F1 is the same as in the original, 0.24 (cut-off 10, stemming applied but not stop-word removal). The difference is that recall and the number of correctly assigned classes increases by 3%, but precision decreases. Thus, depending on the final application, terms extracted in this way could be added to the term list or not.

4.2.1 Implications for application

Enriching the term list with terms extracted using multi-word morpho-syntactic analysis and synonym

<table>
<thead>
<tr>
<th></th>
<th>stemming</th>
<th>stop-words out</th>
<th>all combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>min 1 correct</td>
<td>24479</td>
<td>29639</td>
<td>26039</td>
</tr>
<tr>
<td>min 1 auto</td>
<td>34086</td>
<td>34966</td>
<td>34425</td>
</tr>
<tr>
<td>avg auto/doc</td>
<td>16.79</td>
<td>28.61</td>
<td>18.06</td>
</tr>
<tr>
<td>macroa P</td>
<td>0.11</td>
<td>0.09</td>
<td>0.11</td>
</tr>
<tr>
<td>macroa R</td>
<td>0.54</td>
<td>0.71</td>
<td>0.55</td>
</tr>
<tr>
<td>macroa F1</td>
<td>0.19</td>
<td>0.16</td>
<td>0.18</td>
</tr>
<tr>
<td>microa P</td>
<td>0.07</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>microa R</td>
<td>0.55</td>
<td>0.73</td>
<td>0.59</td>
</tr>
<tr>
<td>macroa F1</td>
<td>0.13</td>
<td>0.11</td>
<td>0.13</td>
</tr>
<tr>
<td>mean F1</td>
<td>0.16</td>
<td>0.13</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Table 7. Performance of the w1 term list enriched with all automatically extracted terms.
acquisition slightly improves recall. At the same time, precision decreases. Thus, enhancing the term list in this way would be appropriate for applications such as focused crawling, when the purpose is to crawl as many documents as possible and precision is less important. To maximize recall, no weights or cut-offs need be applied.

4.3 Term analysis and shortened term lists

In the original term list there were 4,411 distinct terms. In the document collection, 53% of them were found. The average length of the terms found was between one and two words, while the longer ones were less frequently found. Of the terms found in the collection, based on 16% of them correct classes were always found, while based on 43% of them incorrect classes were always found. For a sample of documents containing terms that were shown to always yield incorrect results, we had a male subject expert confirm whether the documents were in the wrong class according to his opinion. For 10 always-incorrect terms with most frequent occurrences, the subject expert looked at 30 randomly selected abstracts containing those terms. Based on his judgments, it was shown that 24 out of those 30 documents were indeed incorrectly classified, but there were also 6 which he deemed to be correct. This is another indication of how problematic it is to evaluate subject classification in general, and automated subject classification in particular. Perhaps one way would be to have a number of subject experts agree on all the possible subjects and classes for every document in a test collection for automated classification; another way could be to evaluate automated classification in context, by end-users.

Based on the term analysis, three new term lists were extracted from the original one, and tested for performance:

1. Containing only those terms that found classes which were always correct (1,308 terms). When cut-off is between 5 and 10, macroaveraged precision reaches 0.89, and microaveraged 0.99, when neither stemming nor stop-words removal are applied. Stemming does not really improve general performance because recall increases only little, by 0.03, while precision decreases by 0.2. However, when using only those 1,308 terms, only 5% of documents are classified. The best mean F1, 0.15, is achieved when stemming and the stop-word removal are used.

2. Containing those terms that found classes which were correct in more instances than they were incorrect (1,924 terms). This list yields best mean F1, 0.38. This value is achieved when stemming is used but no stop-words are removed. There are 65% of documents that are classified, with the average number of classes 1.7. When stemming is not used, precision levels are 0.75 for microaveraged, and 0.79 for macroaveraged.

3. Containing all terms excluding those that found classes which were always incorrect (4,751 terms). The mean F1 is 0.25, when cut-off is 10 and both stop-words removal and stemming are used. The slight improvement in comparison to the original list is due to increase in precision.

4.3.1. Implications for application

Using the same w134_12_1234 term list, apart from by using only weights and cut-offs, precision and F1 are further improved by exclusion of terms that always yield incorrect classes. This setting improves precision without degrading recall, so it should be used in applications when either, or both, are important. The best F1 throughout the whole experiment is achieved when terms that yield incorrect classes in majority of cases are excluded.

5. Conclusion

The study showed that the string-matching algorithm could be enhanced in a number of ways:

1. Weights: adding different weights to the term list based on whether a term is single, phrase or Boolean, which type of class it maps to, and EI term type, improves precision and relevance order of assigned classes, the latter being important for browsing;

2. Cut-offs: selecting as final classes those above a certain cut-off level improves precision and F1;

3. Enhancing the term list with new terms based on morpho-syntactic analysis and synonyms acquisition improves recall;

4. Excluding terms that in most cases gave wrong classes yields best performance in terms of F1, where the improvement is due to increased precision levels.

The best achieved recall is 76%, when the basic term list is enriched with new terms, and precision 79%, when only those terms previously shown to yield
correct classes in the majority of documents are used. Performance of individual classes, measured in precision, is up to 98%. At third and second hierarchical levels mean F1 reaches up to 60%.

These results are comparable to machine-learning algorithms (see, for example, Sebastiani 2002), which require training documents and are collection-dependent. Another benefit of classifying documents into classes of well-developed classification schemes is that they are suitable for subject browsing, unlike automatically-developed controlled vocabularies or home-grown directories often used in document clustering and text categorization (Golub 2006a).

The experiment has also shown that different versions of the algorithm could be implemented so that it best suits the application of the automatically classified document collection. If the application requires high recall, such as, for example, in focused crawling, cut-offs would not be used. Or, if one provides directory-style browsing interface to a collection of automatically classified web pages, web pages could be ranked by relevance based on weights. In such a directory, one might want to limit the number of web pages per class, e.g., assign only the class with highest probability that it is correct, as it is done in the Thunderstone’s web site catalog (Thunderstone 2005).

References


At a time when cataloguing code revision is continuing apace with the consolidation of the *International Standard Bibliographic Description* (ISBD), the drafting of *RDA: Resource Description and Access*, and the development of common principles for an international cataloguing code (International Meeting of Experts on an International Cataloguing Code [IME ICC]), the publication of a guide for cataloguing cultural objects is timely and purposeful. Compiling this data content standard on behalf of the Visual Resources Association, the five editors—with oversight from an advisory board—have divided the guide into three parts. Following a brief introduction outlining the purpose, intended audience, and scope and methodology for the publication, Part One, General Guidelines, explains both what the Cataloging Cultural Objects (CCO) guide is—“a broad document that includes rules for formatting data, suggestions for required information, controlled vocabulary requirements, and display issues” (p. 1)—and is not—“not a metadata element set per se” (p. 1). Part Two, Elements, is further divided into nine chapters dealing with one or more metadata elements, and describing the relationships between and among each element. Part Three, Authorities, discusses what elements to include in building authority records. A Selected Bibliography, Glossary, and Index, respectively, round out the guide.

As the editors note in their introduction, “Standards that guide data structure, data values, and data content form the basis for a set of tools that can lead to good descriptive cataloguing, consistent documentation, shared records, and increased end-user access” (p. xi). The VRA Core Categories, for example, represent a set of metadata elements expressed within an XML structure (data structure). Likewise, the *Art & Architecture Thesaurus* contains sets of terms and relationships, or defined data values. While much effort has been expended on developing both data structures and values, the editors argue, the third leg of the stool, data content, has received less attention. Unlike the library community with its *Anglo-American Cataloging Rules* [sic—though RDA is referenced in the Selected Bibliography], or its archival equivalent, *Describing Archives: A Content Standard* (DACS), those in the domain of cultural heritage responsible for describing and documenting works of art, architecture, cultural artifacts, and their respective images, have not had the benefit of such data content standards. CCO is intended to address (or redress) that gap, emphasizing the exercise of good judgment and cataloguer discretion over the application of “rigid rules” [p. xii], and building on existing standards.

Part One, General Guidelines, sets the foundation. Beginning with the question, “What are you Cataloguing?”, this 41-page section articulates the difference between a work and an image, and continues with what institutions need to consider in determining what kinds of, and how much information to include in, a minimal description for a Work Record—elements subsequently covered in Chapters 1–8 of Part 2—an Image Record—dealt with in Chapter 9 of Part 2—records for a group, collection, or series of cultural objects, and related works, or, “those having an important conceptual relationship to each other” (p. 13). Less familiar, perhaps, to the eyes of those responsible for bibliographic or archival description, is the inclusion of recommendations concerning database design, field structures, database construction, and the purpose of a database—as a cataloguing tool? collection management system? digital asset management system? online catalogue? This latter part, while a useful inclusion, seems somewhat contradictory within a set of guidelines that profess to be “system independent”. Part One concludes with definitions of, and guidelines for, creating and maintaining controlled vocabularies and authority files, respectively. Examples of work records (Figures 1–7), and a work record with two related image records (Figure 8) pro-
vide concrete, visual samples of the issues covered throughout the General Guidelines, and foreshadow the part to follow.

Part Two, Elements, provides (1) definition, context, and terminology, (2) cataloging rules, and (3) guidelines on presentation of data for each of eight broad metadata element types, grouped by purpose, and associated with a work record (e.g., object naming [work type/title]; creator information [creator/creator role]; stylistic, cultural, and chronological information [style/culture/date]; subject; etc.). The ninth chapter, view information elements, addresses how to describe aspects of a work as captured in its surrogate, an image of the work. Each chapter within Part Two concludes with illustrated examples, again, to reinforce concepts and applications discussed relative to a particular element set. Those expecting the inclusion of administrative, structural, and/or technical metadata for creating and managing digital repositories, will be disappointed. The list of elements in Part Two is explicitly restricted to descriptive metadata.

Part Three, Authorities, follows a similar format as Part Two, including discussion and terminology, editorial rules, and presentation of data for (1) personal and corporate name authority, (2) geographic place authority, (3) concept authority, and (4) subject authority. As with Part Two, examples liberally populate the text of each chapter, with specific illustrations of the four types of authority record coming at the end of respective chapters 1–4.

The consistent formatting of chapters within the text, overall, ensures that perspective cataloguers understand the meaning, context, terminology, and application of guidelines for descriptive metadata and authority control. Thus, in its own internal structure, CCO remains true to its stated objective of promoting consistency of interpretation and implementation. Bolded recommendations throughout Part One are, in some instances broad level—“CCO recommends good and versatile database design and consistent cataloging rules” (p. 25)—and in others, appropriately specific—“Because of the complexity of cultural information and the importance of Authority Records, CCO recommends using a relational database” (p. 20). Regardless of their degree of specificity, recommendations provide clear, logical, and principles-based guideposts for both institutions and individual cataloguers, alike. They also provide context for the series of “rules” which follow in Parts Two and Three. The rules, while named as such, and articulated in a prescriptive tone, are discussed and presented throughout in a spirit of “recommended best practice”. This is to allow for individual institutions to “make and enforce” local rules that accommodate their requirements and those of their end-users most effectively and efficiently (p. 2).

This manual will serve as an important tool for museum documentation specialists, visual resources curators, archivists, librarians, or others responsible for providing descriptive metadata and authority control for a variety of cultural objects, including architecture, paintings, sculpture, prints, manuscripts, photographs and other visual media, performance art, archeological sites and artifacts, and different functional objects associated with material culture. While its coverage is impressively wide-ranging, CCO is not intended for natural history or scientific collections.

Cataloging Cultural Objects, in linking the work of cataloguers from different institutional contexts, provides a timely and useful content standard for cross-domain application. It also serves as an effective teaching tool for those who recognize and value, less the location—museum, archive, library—where descriptive metadata are to be assigned, and more the purpose for which they are intended, namely to facilitate access to, and sharing of both records and their corresponding objects. While this reviewer would have appreciated more than a “Selected Bibliography”, and an expanded Glossary (e.g., where is a definition of “format controlled” among “controlled fields”, “controlled list”, and “controlled vocabulary”), the inclusion of additional specialized sources for cataloguing museum collections, and within-chapter references to standard tools for particular metadata elements, are especially foresighted, and commendable. There is mention throughout the text of a “CCO website”. A URL or other link eluded this reviewer, though a Google™ search led to http://vraweb.org/ccoweb/cco/index.html [accessed September 28, 2007].

Overall, Cataloging Cultural Objects with its attending guidelines for descriptive metadata and authority control for “one-of-a-kind cultural objects” should merit a place among the “well-established” data content standards of the library and archival communities that CCO references with obvious regard.

Lynne C. Howarth
Professor
Faculty of Information Studies
University of Toronto
140 St. George Street, Toronto,
Ontario M5S 3G6, Canada
E-mail: lynne.howarth@utoronto.ca

The knowledge and information world we live in can rarely be described from a single coherent and predictable point of view. In the global economy and mass society, an explosion of knowledge sources, different paradigms and information-seeking behaviors, fruition contexts and access devices are overloading our existence with an incredible amount of signals and stimulations, all competing for our limited attention. Taxonomies are often cited as tools to cope with, organize and make sense of this complex and ambiguous environment.

Leveraging an extensive review of literature from a variety of disciplines, as well as a wide range of relevant real-life case studies, *Organising Knowledge* by Patrick Lambe has the great merit of liberating taxonomies from their recurring obscure and limiting definition, making them living, evolving and working tools to manage knowledge within organizations. Primarily written for knowledge and information managers, this book can help a much larger audience of practitioners and students who wish to design, develop and maintain taxonomies for large-scale coordination and organizational effectiveness both within and across societies. Patrick Lambe opens our eyes to the fact that, far from being just a synonym for pure hierarchical trees to improve navigation, findability and information retrieval, taxonomies take multiple forms (from lists, to trees, facets and system maps) and play different roles, ranging from basic information organization to more subtle tasks, such as establishing common ground, overcoming boundaries, discovering new opportunities and helping in sense-making.

Over the course of the book, a number of misconceptions haunting taxonomy work are addressed and carefully dispelled. Taxonomy development is often thought to be an abstract task of analyzing and classifying entities, performed in complete isolation. On the contrary, taxonomies are to a large extent products of users’ perceptions and worldviews, strongly influenced by the pre-existing information infrastructure. They can also be dangerous tools having the potential to reveal and clarify but also to exclude and conceal critical details that can have a large impact on basic business activities such as managing risk, controlling costs, understanding customers and supporting innovation.

If the first part of the book introduces concepts, provides definitions and challenges wrong assumptions about taxonomies and the work of taxonomy-building, the second one takes us step-by-step through a typical project. From here on, insights become part of practicable frameworks that form the basis of a concrete information-management strategy and process so flexible so as to be used in very different organizational environments and scenarios. Starting from the definition of stakeholders, purpose and scope and ending with deployment, validation and governance, a taxonomy-building project is realistically presented as an iterative and fascinating journey over competing needs, changing goals, mixed cues and technical and cognitive constraints.

Beyond introducing fundamental guiding principles and addressing relevant implementation challenges, *Organising Knowledge* provides a large dose of political and pragmatic advice to make your efforts useful in contributing to the overall knowledge and information infrastructure. Taxonomies, much like architect’s blueprints, only represent theory until they are implemented in practice involving real people and real content. As Lambe explains, this step requires crossing over to the other side of the barricade, wearing the user’s shoes and constructing an information neighborhood, designing and populating a metadata framework, solving usability issues and successfully dealing with records management and information architecture concerns.

While each single paragraph of the book is packed with valuable advice and real-life experience, I consider the last chapter to be the most intriguing and ground-breaking one. It’s only here that taxonomists meet folksonomists and ontologists in a fundamental attempt to write a new page on the relative position between old and emerging classification techniques. In a well-balanced and sober analysis that foregoes excessive enthusiasm in favor of more appropriate considerations about content scale, domain maturity, precision and cost, knowledge infrastructure tools are all arrayed from inexpensive and expressive folksonomies on one side, to the smart, formal, machine-readable but expensive world of ontologies on the other. In light of so many different tools, information infrastructure clearly appears more as a complex dynamic ecosystem than a static overly designed environment. Such a variety of tasks, perspectives, work activities and paradigms calls for a resilient, adaptive and flexible knowledge environment with a minimum of standardization and uniformity. The right mix of tools and approaches can only be deter-
mined case by case, by carefully considering the particular objectives and requirements of the organization while aiming to maximize its overall performance and effectiveness.

Starting from the history of taxonomy-building and ending with the emerging trends in Web technologies, artificial intelligence and social computing, Organising Knowledge is thus both a guiding tool and inspirational reading, not only about taxonomies, but also about effectiveness, collaboration and finding middle ground: exactly the right principles to make your intranet, portal or document management tool a rich, evolving and long-lasting ecosystem.

Emanuele Quintarelli
E-mail: emanuele.quintarelli@gmail.com
ISKO News

Edited by Hanne Albrechtsen
Communications Editor

ISKO’s Nordic chapter was founded November 8th 2007 in Copenhagen and covers Sweden, Denmark, Norway, Finland, Iceland & the Faroe Islands. We will eventually approach the Baltic countries and determine their interest in the project. Its first board constituted itself with Mikkel Christoffersen (DK) as chairman and as board members professor Birger Hjørland (DK), Hanne Albrechtsen (DK) and Per Nyström (SWE). 22 people have so far expressed interest in the chapter, but before we see who pays the membership fee for 2008, we do not yet know how many members we will be.

We will establish a web presence soon, and the plan is to hold a conference in odd years with the first one being 2009 in Sweden. The theme of the first conference will probably be whether there is a Nordic school of thought in knowledge organisation. We hope the chapter will bring together the Nordic researchers in KO and facilitate more communication and exchange of ideas as well as cooperation and general awareness of ongoing research and the involved parties therein.

Mikkel Christoffersen, ph.d.-stipendiat
Danmarks Biblioteksskole / Royal School of Library and Information Science
Knowledge Organization Literature

Ia C. McIlwaine: Literature Editor

Assisted by: Marie Baliková, Victoria Frâncu, Claudio Gnoli, Ágnes Barátné Hajdu, John McIlwaine, Gerhard Riesthuis, Aida Slavic, Rosa San Segundo, Alenka Sauperl, Nancy Williamson.

Without their assistance the task would not be possible, and their help is greatly appreciated, as would be contributions from any other willing person.

ICM

0 Form division

02 Literature Reviews in Knowledge Organization


04 Universal Classification Systems


06 Conference Reports and Proceedings

07 Textbooks (whole field)

0491 07.23
Book reviews by:

0494 07.3
Book reviews by:

1 Theoretical Foundations and general Problems

11 Order and Knowledge Organization

0497 111

0498 111

12 Conceptology in Knowledge Organization

0499 122

0500 122

0501 122
2 Classification Systems and Thesauri, Structure and Construction

21 General Problems of Classification Systems and Thesauri


22 Structure and elements of CS & T


0525 Shimada, M. - The revision of the National Diet Library List of Subject Headings (NDLSH), and its future (Lang.: jap). – In: Journal of Information Science and Technology Association (Joho no Kagaku to Gijutsu), 57(2007)2, p.73-78.

23 Construction of Classification Systems and Thesauri


24 Relationships


25 Numerical Taxonomy


26 Notation. Codes


29 Evaluation of C S & T


34 Classing and Indexing


Li, T., Zhu, S., Ogihara, M. - Hierarchical document classification using automatically generated hierarchy (Lang.: eng).
35 Manual and Automatic Order Techniques


36 Coding


39 Evaluation of Classing and Indexing


4 On Universal Classification Systems and Thesauri


43 On the Dewey Decimal Classification


<table>
<thead>
<tr>
<th>Code</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
<td>563</td>
</tr>
<tr>
<td>44</td>
<td>564</td>
</tr>
<tr>
<td>44</td>
<td>565</td>
</tr>
<tr>
<td>44</td>
<td>566</td>
</tr>
<tr>
<td>48</td>
<td>567</td>
</tr>
<tr>
<td>48</td>
<td>568</td>
</tr>
<tr>
<td>6</td>
<td>569</td>
</tr>
<tr>
<td>6</td>
<td>570</td>
</tr>
<tr>
<td>44</td>
<td>571</td>
</tr>
<tr>
<td>44</td>
<td>572</td>
</tr>
<tr>
<td>44</td>
<td>573</td>
</tr>
<tr>
<td>44</td>
<td>574</td>
</tr>
<tr>
<td>66</td>
<td>575</td>
</tr>
<tr>
<td>66</td>
<td>576</td>
</tr>
<tr>
<td>66</td>
<td>577</td>
</tr>
<tr>
<td>66</td>
<td>578</td>
</tr>
<tr>
<td>66</td>
<td>579</td>
</tr>
</tbody>
</table>

#### 44 On the Library of Congress Classification and Library of Congress Subject Headings

- **563**  

- **564**  

- **565**  

#### 48 On other Universal Classification Systems and Thesauri

- **566**  

- **567**  
  *The main purpose and the basic rules of developing "ExpressFinder/Thesaurus"* (Lang.: jap). – In: Journal of Information Science and Technology Association (Joho no Kagaku to Gijutsu), 57(2007)2, p.84-88.

#### 6 On Special Subjects Classifications and Thesauri

- **568**  

- **569**  

#### 62 On C & T in Physics, Chemistry, Electronics, Energy

- **570**  

- **571**  

#### 65 On C & T in Human Biology, Medicine, Psychology, Education, Labour, Sports, Household

- **572**  

- **573**  

#### 66 On C & T in Sociology, Politics, Social Policy, Law, Area Planning, Military Science, History

- **574**  
60 On C S & T in Science of Science, Information Science, Computer Science, Communication Science, Semiotics


7 Knowledge Representation by Language and Terminology

71 General Problems of Natural Language in Relation to Knowledge Organization


73 Automatic Language Processing


74 Grammar Problems


75 On-Line Retrieval Systems and Technologies

751 General and Theoretical Problems


0595

0596

0597

752 Dialogue systems. Interactive Catalogues
0598

0599

0600

0601

753 Online Access, Queries, Free Text Searching
0602

0603

757 Expert Systems in Searching. Search Engines
0604


0605

0606

0607

0608

759 Evaluation of On-Line Information Retrieval Systems and Techniques
0609

0610

76 Lexicon/Dictionary Problems
0611

0612

77 Problems of Terminology
0613


8 Applied Classing and Indexing

81 General Problems, Catalogues, Guidelines, Rules, Indexes

0631 81

82 Data Classing and Indexing

0632 82-92;43

83 Titled classing and Indexing. Derived Indexing, Folksonomy

0633 831

0634 835

0635 835

0636 835

84 Primary Literature Classification and Indexing

0637 842

0638 842

0639 842

0640 842

0641 844;918

0642 846;88-46

85 (Back of the) Book Classification and Indexing

0643 854

0644 854

86 Secondary Literature Classification and Indexing

0645 864

0646 864

0647 864

0648 864
Sawyer, S., Huang, H. – Conceptualizing information, technology and people: comparing information science and in-


87 Classification and Indexing of Non-Book Materials


Book reviews by:


88 Classification and Indexing in Subject Fields


9 Knowledge Organization Environment

91 Professional and Organisational Problems in General and in Institutions


Book reviews by:


98 User Studies


Personal Author Index

Abbas, J. 0634
Acosta, S. 0669
Agosti, M. 0548
Ahn, H. 0583
Amaral, C. 0622
Andrews, J.E. 0482
Angrosh, M. L. 0505
Arazy, O. 0535
Arenčíková-Jorge, R. 0572
Arnott, G. 0675
Baca, M. 0664
Baliková, M. 0623
Balkhatir, M. 0662
Barátné Hajdu, Á. 0499, 0500
Basile, P. 0536
Bawden, D. 0691
Bilodeau, B. 0567
Bovbel, E. 0654
Braun, T. 0692
Buzza, P. 0522
Burton, W.C. 0579
Calhoun, K. 0670
Carenfil, T. 0553
Cathey, R.J. 0529
Chang, C.-C. 0540
Chapman, A. 0671
Chapman, J.W. 0665
Charhad, M. 0663
Chatterjee, K. 0653
Chen, G. 0650
Chen, S.-C. 0653
Chen, Y.-H. 0615
Chen, Y.-L. 0609
Chen Yintao 0632
Chen, Z. 0507, 0617
Cheng, L.-C. 0609
Cheng, Y.-L. 0609
Cheung, K. 0526
Chiang, Huei-Min 0503
Chihiaia, L. 0688, 0689
Chipaux, E. 0688
Choi, Key-Sun 0639
Chou, C.-H. 0615
Chun-Yao Huang 0696
Clavel, P. 0629
Cole, C. 0482, 0697
Conduit, N. 0658
Conway, C.N. 0495
Cordeiro, I. M. 0554
Costa, V.S. 0537
Covaci, M. 0688
Cuvillier, J. 0641
Czerniejewski, B. 0624
Dubney, D. 0578
Dahlgren, I. 0645
Dalbin, S. 0510, 0511
D’Ambrosio, D.M. 0176
Daniel, H.-D. 0645
Danskia, A. 0681
De Campos, L.M. 0542
Dehuri, S. 0588
Dina, Y. 0580
Dmitry, M. 0564
Doucet, A. 0549
Druin, A. 0591
Dunlay, D.M. 0530
Du Preez, M. 0497
Efron, M. 0552
Erdeléi, S. 0698
Faber, P. 0614
Ferraioli, L. 0678
Perro, N. 0458
Fielding, E. 0659
Flehnert, C. 0557
Fong, A.C.M. 0607
Fourie, I. 0603
Francis, E. 0635
Fráncu, V. 0555
Frosch, P. 0666
Fu, Y. 0590
Fu Lee Wang 0616
Fugmann, R. 0643
Gagnon, G. 0569
Galvez, C. 0690
Gery, M. 0550
Giunchiglia, F. 0544
Given, L.M. 0698
Goh, D.H. 0587
Gómez González-Jover, A. 0611
Gopal, T.V. 0522
Goyal, O.P. 0598
Han, C.-C. 0615
Han, I. 0583
Han, L. 0605
Harai, N. 0630
Harcourt, K. 0620
Harper, C.A., 0533
Haslinger, I. 0631
Hatfield, A.J. 0679
Haus, G. 0620
He, W. 0699
Hert, C.A. 0673
Heuwing, B. 0625
Hickey, L. 0579
Hidderley, R. 0682
Hillson, B. 0565
Hjørland, B. 0512
Hoover, L. L. 0642
Horo-witz, E. 0546
Hsiieh-Lee, I. 0655
Hsu, C.-F. 0686
Hu, J. 0589
Hu, Y. C. 0595
Hu Yuefang 0632
Huang, Chun-Yao 0696
Huang, H. 0648
Hui, S.C. 0627
Hunter, K. 0518
Hunter, J. 0526
Hutchinson, H.B. 0591
Intner, S.S. 0496, 0674
Isaac, A. 0684
Jansen, J. 0584
Janceck, P. 0597
Jimenez, A. G. 0515
Johnston, B.H. 0569
Kang, In-Su 0638
Kapoor, K. 0598
Karamuťuoglu, M. 0501
Karypis, G. 0606
Kasten, J. 0506
Ke, W. 0690
Kelley, S.D. 0679
Kelsey, R.J. 0633
Khairy, I. 0558
Kharkevich, U. 0544
Khoo, C.S.G. 0587
Kim, Jae-Ho 0639
Kim, K. 0583
Kim, S. 0532
Kishida, K. 0513
Kofmanov, L. 0680
Kousha, K. 0646
Kovac, T. 0556
Kuntz, B. 0565
Kunze, H. 0643
Kuramoto, M. 0606
La Barre, K. 0521
Lacoste, C. 0573
Laurent, D. 0622
Lazar, J. 0596
Lazarinis, F. 0585
Laing, S.S. 0674
Lehtonen, M. 0549
Leman, S. 0667
Levinson, D. 0576
L’Homme, M.C. 0613
Li, T. 0545
Li, Y. 0540
Li, Y.-C. 0540
Liao Jianxin 0596
Lin, Wen-Yau C. 0519
Lioma, C. 0543
Lisek, S. 0626
Loh, H. T. 0541
Lopes, R. 0537
López-Huertas, M.J. 0577
Lorringer, L. 0693
Lu, K. 0507
Madalli, P. 0516
Mall, R. 0588
Mandl, T. 0622, 0625
Manevitz, L. 0504
Mangan, E. 0663
Mao, C.-C.A. 0686
Martí-Lahuerta, Y. 0572
Matthews, D. 0644
Mazzocchi, F. 0621
Ménard, E. 0627
Michon, J. 0677
Miksa, F. 0479
Miksa, S.D. 0520
Miller, D.P. 0494
Mitchell, J.S. 0562, 0619
Modeste, J. 0580
Montgomery, P. 0559
Moore, J.L. 0699
Merk, K. 0636
Mortchez-Bouveret, M. 0528
Mostafa, J. 0590
Moya-Anegón, F. 0690
Muñ-Chuyen Tang 0574
Munk, T.B. 0636
Naumis-Peña, C. 0581
Nero, L. M. 0619
Neumann, A. 0592
Nielsen, S. 0612
Niemi, T. 0584
Nissen, M.E. 0497
Ogihara, M. 0545
Ojala, M. 0524
O’Leary, M. 0601
Orphan, S. 0566
Ou, S. 0587
Ouniis, I. 0543
Panici, A. 0508, 0687
Peng, D. 0538
Pestov, V. 0575
Petgnet, D. 0560
Petrou, A.D. 0676
Pinto, A. 0620
Ponnusamy, R. 0502
Popa, M. 0689
Poulter, A. 0694
Quensel, O. 0635
Quinn, S. 0492
Raban, D.R. 0610
Rafferty, P. 0658, 0682
Ramírez, I de T. 0577
Rédy, G. 0592
Richman, D. 0561
Rorissa, A. 0660
Rousseau, R. 0467
Rowley, J. 0527
Ru, Y. 0546
Ryoo, J. 0593
Sagonas, K. 0537
Saiedian, H. 0593
Salo, J. 0498
Samantray, S. 0539
Satija, M.P. 0531
Sato, H. 0568
Sato, H. 0472
Sato, H. 0568
Sato, H. 0568
Sato, H. 0568
<table>
<thead>
<tr>
<th>Name</th>
<th>Page Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sawyer, S.</td>
<td>0648</td>
</tr>
<tr>
<td>Schiff, A.</td>
<td>0564</td>
</tr>
<tr>
<td>Schneider, J.W.</td>
<td>0649, 0652</td>
</tr>
<tr>
<td>Seabrook, N.</td>
<td>0657</td>
</tr>
<tr>
<td>Shawe-Taylor, J.</td>
<td>0640</td>
</tr>
<tr>
<td>Shen, J.-J.</td>
<td>0540</td>
</tr>
<tr>
<td>Shi, Xiaodong</td>
<td>0608</td>
</tr>
<tr>
<td>Shimada, M.</td>
<td>0525</td>
</tr>
<tr>
<td>Slavic, A.</td>
<td>0582</td>
</tr>
<tr>
<td>Smiraglia, R.</td>
<td>0490</td>
</tr>
<tr>
<td>Smith, S.</td>
<td>0557</td>
</tr>
<tr>
<td>Song, D</td>
<td>0547</td>
</tr>
<tr>
<td>Soonja Lee Koh, G.</td>
<td>0523</td>
</tr>
<tr>
<td>Spink, A.</td>
<td>0482</td>
</tr>
<tr>
<td>Spotti Lopes Fujita, M.</td>
<td>0695</td>
</tr>
<tr>
<td>Stirling, D. A.</td>
<td>0571</td>
</tr>
<tr>
<td>Stojmirović, A</td>
<td>0575</td>
</tr>
<tr>
<td>Strossa, P.</td>
<td>0586, 0628</td>
</tr>
<tr>
<td>Strotgen, R.</td>
<td>0625</td>
</tr>
<tr>
<td>Sukula, S. K.</td>
<td>0534</td>
</tr>
<tr>
<td>Sutó, Z.</td>
<td>0592</td>
</tr>
<tr>
<td>Tang, Muh-Chyun</td>
<td>0574</td>
</tr>
<tr>
<td>Taylor, A.G.</td>
<td>0494</td>
</tr>
<tr>
<td>Tennis, J.T.</td>
<td>0551</td>
</tr>
<tr>
<td>Tenopir, C.</td>
<td>0700</td>
</tr>
<tr>
<td>Thelwall, M.</td>
<td>0646</td>
</tr>
<tr>
<td>Tho, Q.T.</td>
<td>0607</td>
</tr>
<tr>
<td>Tiberi, M.</td>
<td>0621</td>
</tr>
<tr>
<td>Tillett, B.</td>
<td>0533</td>
</tr>
<tr>
<td>Trickey, K.V.</td>
<td>0493</td>
</tr>
<tr>
<td>Uddin, M.N.</td>
<td>0597</td>
</tr>
<tr>
<td>Ungváry, R.</td>
<td>0517</td>
</tr>
<tr>
<td>Uralman, N. H.</td>
<td>0668</td>
</tr>
<tr>
<td>Urs, S. R.</td>
<td>0505</td>
</tr>
<tr>
<td>Van der Linden, H.</td>
<td>M M. 0509</td>
</tr>
<tr>
<td>Van Otegem, M.</td>
<td>0631</td>
</tr>
<tr>
<td>Vivian, I.</td>
<td>0600</td>
</tr>
<tr>
<td>Wackern, M.</td>
<td>0600</td>
</tr>
<tr>
<td>Wang, Fu Lee</td>
<td>0616</td>
</tr>
<tr>
<td>Wang, Tai-Yue</td>
<td>0503</td>
</tr>
<tr>
<td>Weihs, J.</td>
<td>0562, 0674</td>
</tr>
<tr>
<td>Wells, D.</td>
<td>0599</td>
</tr>
<tr>
<td>Whited, M.</td>
<td>0618</td>
</tr>
<tr>
<td>Williamson, N.J.</td>
<td>0594</td>
</tr>
<tr>
<td>Wolley, I.</td>
<td>0600</td>
</tr>
<tr>
<td>Woo, C.</td>
<td>0535</td>
</tr>
<tr>
<td>Xiaodong Shi</td>
<td>0608</td>
</tr>
<tr>
<td>Xie, L.</td>
<td>0605</td>
</tr>
<tr>
<td>Yang, C.</td>
<td>0516</td>
</tr>
<tr>
<td>Yang, C.C.</td>
<td>0608</td>
</tr>
<tr>
<td>Yee, M. M.</td>
<td>0683</td>
</tr>
<tr>
<td>Yousaf, M.</td>
<td>0504</td>
</tr>
<tr>
<td>Yuefang, Hu</td>
<td>0632</td>
</tr>
<tr>
<td>Zaihrayeu, I.</td>
<td>0544</td>
</tr>
<tr>
<td>Zanotto, E. D.</td>
<td>0651</td>
</tr>
<tr>
<td>Zhan, J.</td>
<td>0541</td>
</tr>
<tr>
<td>Zhang, R.</td>
<td>0595</td>
</tr>
<tr>
<td>Zhu, S.</td>
<td>0545</td>
</tr>
<tr>
<td>Zhu Xiaomin</td>
<td>0596</td>
</tr>
</tbody>
</table>
In Zeiten schrumpfender Zuschüsse aus der öffentlichen Hand müssen Wirtschaft und Kultur zusammenarbeiten, um Kultur als für beide Partner wichtige Säule am Leben zu halten.

Wie findet man Sponsoren? Und wo? Wie spricht man sie an? Wie findet man zu einer beidseitig fruchtbaren Partnerschaft?

“Sponsoren gewinnen - aber wie?”
Trainings für Kulturinstitutionen und Künstler
These essays by noted Area Studies specialists at a number of US research libraries serve as a practical and theoretical guide to university and college administrators, library directors and heads of collection development, as well as selection practitioners who work to create foreign-language collections for research libraries. The volume constitutes a general introduction for new practitioners and even the most experienced Area Studies librarians will find useful practical advice for reviewing and refining their existing collecting practices. Coverage includes the Middle East, East Asia, Latin America, Southeast Asia, Africa, and the Romance language areas of Europe, as well as the German/Nordic/Netherlandic countries. Each essay presents the Area Studies topic in question from an historical perspective and provides background on its present status and anticipated future development. Special emphasis is placed on the techniques of both print and digital collecting and on the assessment methods by which collection strengths and future needs are determined. Guidelines for expenditures for both collections and collateral activities such as providing access and preservation are provided, and contributors also supply extensive documentation for the burgeoning array of online digital resources which have emerged in the past decade. The volume editors, Dan C. Hazen (Harvard) and James H. Spohrer (University of California, Berkeley), also provide a general introduction to the topic and a detailed summary of current cooperative activities in Area Studies collecting.