Keynote Panel: ISKO Fifth International Conference 1998, Lille, Paris

Edited by Rebecca Green

College of Library and Information Services, University of Maryland

Panel Members: Ia McIlwaine, A. Neelameghan, Michèle Hudon, Christian Fluhr, Joan Mitchell, Carol Bean, and Rebecca Green

Rebecca Green is on the faculty of the College of Library and Information Services at the University of Maryland, College Park, MD, where she teaches in the knowledge organization area. Her special emphases include database design and cognitive linguistics. Dr. Green was the program chair of the 4th International ISKO Conference held in 1996 in Washington, DC.

Green, R. (Ed.). (1998). Keynote Panel: ISKO's Fifth International Conference: Lille, France. *Knowledge Organization, 25(4),* 143--161.

ABSTRACT: ISKO's Fifth International Conference was held August 25-29, 1998 in Lille, France. The conference opened with a panel (Ia McIlwaine, A. Neelameghan, Michèle Hudon, Christian Fluhr, Joan Mitchell, and Carol Bean; moderator, Rebecca Green) who addressed the general theme of the conference, Structures and Relations in Knowledge Organization. Each panelist was given a brief period in which to reflect on issues in one or more of three areas falling within the conference theme: (1) The role of hierarchical relationships in knowledge organization; (2) Relationships in multilingual, multicultural, and multidisciplinary contexts; and (3) Relationships in online retrieval

ISKO's Fifth International Conference was held August 25-29, 1998 in Lille, France. The conference opened with a panel (Ia McIlwaine, A. Neelameghan, Michèle Hudon, Christian Fluhr, Joan Mitchell, and Carol Bean; moderator, Rebecca Green) who addressed the general theme of the conference, Structures and Relations in Knowledge Organization. Each panelist was given a brief period in which to reflect on issues in one or more of three areas falling within the conference theme:

(1) The role of hierarchical relationships in knowledge organization
(2) Relationships in multilingual, multicultural, and multidisciplinary contexts
(3) Relationships in online retrieval

In the time remaining after the panelists presented their personal remarks, there was limited opportunity for questions from the floor.

The areas were presented to the panelists as a set of questions (it should be noted, however, that the panelists were not asked to address the questions *per se*, nor were the areas implied by the questions considered to exhaust the conference theme):

1. Hierarchy has been the dominant structuring mechanism in knowledge organization. Can the strengths of hierarchy be achieved/approached with other structuring devices? If so, sketch the use of alternative structures. Are there pitfalls associated with an over-reliance on hierarchy? How well can hierarchy co-exist with other structuring mechanisms?
2. Are relationships in knowledge organization necessarily constrained by such contexts as language, culture, and discipline? Is there a universal set of relationship types applicable across all these contexts? How much variation exists in how relationships are envisaged across nominally-equivalent concepts (i.e., people use the same word or phrase, but the underlying concepts are not exactly the same)? How can we build integrated knowledge organization schemes that reflect a multiplicity of relational views?

3. Most Internet search engines do not take explicit account of relationships, but instead rely on stem/word/phrase occurrences in text. The problems inherent in this approach are well-known. What role should relationships play in retrieval in the online environment? Is the incorporation of a relational approach to retrieval feasible, given the volume and diversity of materials online? How could we evaluate the impact of incorporating a relational approach to online retrieval?

The presentation order of the panelists was arranged so as to group their remarks by the area or areas they had chosen to address. Interest in relationships across language, culture, and discipline boundaries proved to be the area of the highest degree of interest, and consequently that area was addressed by the first several of the panelists. Focus on the other two areas is found in the remarks of the later panelists.

Professor McIlwaine is Director, School of Library, Archive and Information Studies, University College London and Editor in Chief of the Universal Decimal Classification (UDC).

Some Problems of Context and Terminology

Ia C. McIlwaine

Three questions have been posed from which we have been asked to select a theme that will set the scene for the discussions that we are to have in the next few days. The one that I find most interesting is the second, which asks whether relationships in knowledge organization are necessarily constrained by such contexts as language, culture and discipline. Some of the answers to this question also relate to the first one which is concerned with hierarchy, but take the argument slightly further. The really problematical relationships for information retrieval are not those that are hierarchical, since a hierarchy in the abstract is a readily understood concept. It is how that hierarchy is constructed that is more subjective and this immediately leads into the second question of context and a mutual understanding of what is meant. It is therefore on the problems of context and terminology that I would like to concentrate.

One of the major problems facing designers of information systems for subject retrieval, especially if the system is aiming at some kind of universal usage, is that of context. This is particularly true of systems that rely totally on words, rather than a systematic structure of some kind, as the principal basis for retrieval. Single terms are frequently meaningless, and even more frequently have a multiplicity of meanings when they occur alone, and only take on a proper meaning when given a context, e.g., pest control, self control, the controls of an aeroplane, control of the state, etc., etc. This problem has to be faced whatever type of retrieval tool is being designed, and creators of classification schemes, of thesauri or of other types of indexing language have tried a range of different approaches to combat it. I would suggest that the most successful and probably the most expensive solution is the use of a totally structured system, ideally a classification scheme, but even that requires a willingness to submit to certain constraints and to adhere to certain standards that may not be entirely satisfactory to the user, or understood by him or her.

All schemes of classification are constructed within a given framework and make certain common assumptions, whether they are the DDC, LC or the BBK. The "bibliographic imperialism" of standards created in the West is put up with by others in the interests of economy, but it is by no means received without considerable discomfort. So, the contexts of time and culture are powerful factors. Another element that should not be overlooked is that of the specialism or discipline from which the enquirer comes. The same term may be sought with very different intent by, for example, a geologist or an archaeologist, or a chemist and an electrician-ionization will mean one thing to a chemist and something quite different to an engineer.

Even the most rudimentary of indexing systems based on keywords in titles of books or articles (e.g., KWIC or KWOC) depend upon a context, though of a rather different kind. When the terms in a title are permuted the context helps the user to locate items that are relevant and to discard those that are not. The success of such a system, however, depends upon the author of the article in question having given it a meaningful title. But how can one cope with context in retrieval, especially in the humanities where such an approach is less common and where titles are frequently devoid of meaning until the item in question has actually been read? Titles taken out of context are frequently incomprehensible, or may be meaningless without some prior knowledge on the part of the reader. A particularly acute instance is when they are based on a quotation that needs to be recognized for the full impact to be conveyed, e.g., "Devices and desires" or the "Wooden O".

Context cannot be looked at without also considering the terminology used, because for a full understanding they are interdependent. The meaning of terms, es-

pecially in a language such as English, which lacks any "authority control" in the way that French has in the form of the Académie or German through the Duden, depends heavily upon an understanding that is developed through education, through upbringing and environment, through reading, through the media, especially the television, and through everyday encounter. And English is very frequently assumed to be universally understandable, so this lack of authority control is problematical. Terminology used may have strong dependence upon an understanding of a given context in order to make an impact. Advertisements are one very clear demonstration of this – "the world is orange", "which washing machine is greener?"

The problem of context is one to which a range of solutions, all only partially successful, has been applied. The most obvious in an online situation is the use of Boolean operators and techniques such as stemming and truncation. But in order to work satisfactorily this must be linked to a controlled vocabulary, and the success of this is limited, and can lead to some very strange results:

Rape *(May Subd. Geog.)*
Rape (Islamic Law) *(May Subd. Geog.)*
Rape (Plant) *(May Subd. Geog.)*
Rape (Roman Law)
Rape in Marriage *(May Subd. Geog.)*
Rape of the Sabine Women (Legend)

All these measures are familiar to us in thesauri or subject headings lists. But the dependence upon a controlled vocabulary throws great onus on:

(a) The user, to find the correct term;
(b) The designer of the system, to sort out the synonyms coherently and comprehensively, possibly by means of presenting them transparently and allowing the user to sort them for him/herself; and
(c) Very often, the assumption that the user speaks English, or at least has sufficient grasp of the language to use the system properly.

This last is a particular problem, since many English-speaking countries use different terms for the same thing and even more problematically, the same term has a different meaning depending upon which side of the Atlantic one comes from, let alone, Australia, New Zealand, South Africa or anywhere else in the British Commonwealth. "Correct" English is actually most likely to be found in the English-speaking Caribbean or India. The problem is further enhanced by the fact that one does not know in what "tradition" a non-native speaker has been taught English.

There is a very real constraint imposed by language, culture and discipline, and I think this is virtually insuperable because, unless one is a native speaker, and not always even then, it is not possible fully to understand the meaning of a piece of writing, aside from the strictly scientific or technical. Science and technology probably present the least problems for retrieval and are the areas in which the most work has been undertaken to resolve the barriers in communication. They are also fields where scholarly discourse tends to be in English, and there is mutual understanding of what is meant.

It is not only in the Humanities where "woolly" terminology can cause difficulties. A proper understanding of terminology in the social sciences is heavily dependent upon an understanding of context, particularly because in these disciplines there is a strong tendency to borrow terminology and use words in a different sense from that in which they were originally used; even in science, and above all in medicine, there is a problem with words that are a mixture of Greek and Latin and do not always convey the same meaning as the words in the original language, thereby confusing someone who understands Greek or Latin and so may wrongly assume they have understood the concept under discussion. (Philologically speaking, "metadata" is a completely meaningless, hybrid word that we see frequently nowadays). When one reaches the disciplines that depend upon the imagination, and especially the realms of literature where metaphor, for example, is a common conceit, the difficulties are even more obvious.

Both terminology and context become vital factors when one is trying to devise or revise a classification designed to cover the whole universe of knowledge. One of the first people to attempt to grapple with the problem of "concretes" was Brown with his Categorical Table. It is not, however, really helpful to locate roses in Botany and then express a range of contexts such as heraldry, flower arrangement, roses as an architectural decoration, "Farewell, England's rose" etc., through the use of an auxiliary table. Neither is it helpful (except to the specialist, who will not want everything the database holds on a given topic) to enumerate with different notations terms such as the rose or the pea in a number of different places, such as Botany, Gardening, Agriculture, Cookery, etc., as many classification schemes, including the UDC, do. But for successful retrieval, it is essential that the context is conveyed by some means or other, and the better the system is at expressing context, the more useful it becomes.

The subsequent question seeks to discover whether there is a universal set of relationships that is applicable across all these contexts? Many have attempted to find

one; e.g., Farradane, or Austin with PRECIS. All have failed to gain currency, largely for economic rather than intellectual reasons. They require skilled indexers for their implementation, they have an elaborate structure and they are complicated and therefore expensive. But I wonder whether these systems have not received widespread acceptance also, because many indexers are brought up to place too much reliance on the index to the classification scheme, a situation that is much encouraged by Dewey and his relative index, and has conditioned many generations of indexers to bad habits. The lack of skilled indexers, and the fact that it is not possible to place total reliance on any two indexers, however well trained, to produce the same solutions, was amply demonstrated in the Cranfield experiments in the 1960s and their findings have yet to be refuted. Since the 1960s the volume of material has increased dramatically, with the result that sensible indexing is virtually precluded. Much indexing nowadays is done in a very ad hoc fashion and relies on retrieval systems, through devices such as weighting and proximity, to supply the context in a quasi-post coordinate way at the point of retrieval. To revert to Dewey, the 17th edition of his scheme was equipped with an index that did not supply the "right" answer, but required the user to think, and was universally considered to be a disaster as a consequence.

There are really two sets of obstacles present in this set of questions The first is one of definition (terminology) and the second is one of relationships (context). A well-designed system for expressing universal relationships is not an impossible goal, in the abstract. The standards for thesaurus construction spell them out quite clearly, though the most problematical area lies in the RTs, not the BTs and NTs which can be solved through an understanding of hierarchical principles and a clear display of the selected hierarchies. The limitation here is the one referred to earlier, dependence upon an understanding and acceptance of how those hierarchies have been constructed. But the designation of related terms is far more subjective and relies heavily upon the connection of ideas in the mind, which gets us back to such matters as context and culture.

A system for expressing relationships must depend upon a linking of clearly defined concepts, and I suspect that the initial problem is that of definition rather than relationship. The larger the database that is being handled, the greater the problem becomes and when it reaches Internet dimensions it appears to be in the realms of the impossible. The sheer scale of possible retrieved sets would seem to indicate that even good schemes for sorting and providing the appropriate context would still leave the actual users with many thousands of true hits, for which they would then need some internal form of relevance ranking. And most searches on this scale are probably one-shot, with no refinement.

To be successful, any system must make assumptions, one of which must inevitably be outlook or viewpoint – all the general schemes of classification do, and this is probably why the most successful ones are those designed specifically for one institution – to take an extreme example, the Library of Congress Classification makes much better sense in the context and historical circumstances of that Library, than it does in, say the National Library of Wales or the London School of Economics. One possible method might be to build into an online system a large and accepted dictionary – for English, the Oxford English Dictionary in its full text, so that it is possible to select terms and to see how they are used in specific contexts through the use of the quotations supplied to illustrate the various meanings of the term. If the searcher could first select the definition that best represents what he is seeking, and this could then be linked to a systematic display that flagged up the various relationships in their correct context it might be possible to make use of a system of relationships to retrieve the required information, along the lines of those well known already in the field through the work of such people as Farradane or Austin and others. But this would be an extremely slow and cumbersome solution. It is also not entirely removed from using a classification scheme which, again, will put the term in its context. The difference between using an authority based on a standard dictionary underpinned by a collection of quotations and using one or more of the standard classification schemes, would be that the user has to make the decision of where to turn next, and perhaps make several choices, rather than rely on someone else's preconceived idea of how a classification should be constructed. Perhaps the greatest problem of all is that we are seeking instant solutions and the requirement for speed inevitably leads to the cutting of corners – the more thorough the underlying system, the slower the retrieval of information and the more expensive the task of indexing that information becomes.

Professor Neelameghan is Honorary Visiting Professor, Documentation Research and Training Center (DRTC), Indian Statistical Institute (ISI), Bangalore, India.

Lateral Relations and Links in Multicultural, Multimedia Databases in the Spiritual and Religious Domains: Some Observations

A. Neelameghan

1. Introduction

Emerging information and communications technologies, powerful software for networking, and hypertext linking enable information seekers rapid access to and retrieval of vast numbers of a variety of information records (texts, directories, tables, graphics, images, sound) from multiple databases, located even at global distances. On the Internet, for instance, even on topics such as Vedanta and Christian Mysticism, the first retrieval hits some 2000 items and if the link >More like this' is used for search one gets an additional 100,000 hits on each topic! In many cases, it is possible to download selected records onto a PC database. Yet the sheer volume and variety of records retrieved make it difficult to select the really pertinent items for use. As is well known, relevance of information depends on a number of factors. These relate to the original source materials, the way they were indexed for retrieval, the vocabulary control tools used, if any, at the time of data entry, capabilities of the search engine, user interface, and factors relating to the information seeker including his/her prior experience in using online and other databases. With multiple and cross-cultural databases the difficulties may be more pronounced than with a single database for a well-defined field. Standardization and consistency in the rendering of names or codes used as surrogates, of persons, corporate bodies, subjects, etc., in the databases can improve the efficiency of retrieval and pertinence of the information retrieved.

The observations presented here relate to our experiences in dealing with user interfaces, vocabulary control, thesaurus construction, classification, and related issues while designing, developing and using multimedia, multicultural and mostly non-bibliographic databases in the spiritual and religious domains. There are three versions: ported on to the Internet [accessible at addresses http:// 144.16.72.175/~om/ or http://ukko.grainger.uiuc.edu/omasp/ (without sound)], CD-ROM [contact: raja@ncsi.iisc.ernet.in] and stand alone PC, which constitute the OM Information Service (OMIS). No definitive solutions are offered; only what is being done at present are mentioned. Specifically Non-hierarchic Associative Relations – hereafter called Lateral Relations (LR)Camong concepts will be considered. Some twenty-five years ago I reported on some thirty types of LR and how they were handled in S.R. Ranganathan's Colon Classification and how to use the LR in generating RT's in thesaurus building.

2. OM Information Service (OMIS) Databases

The technical details of the OMIS are described elsewhere [1]. But some minimum information about the databases, indexing and search facilities of OMIS would help in placing the issues in the proper context. OMIS is multimedia (text, sound and image) and consists of three databases that are inter-linked (hypertext links). The three databases are:

1. OM: A database of extracts (currently about 18000) from the sayings, discourses, poems, writings, etc. of religious leaders, mystics, saints, seers, prophets and scholars and from religious texts and epics (the Bible, the Koran, Ramayana, Bhagavad Gita), together called Sources (about 900), from 3000 BCE to the present, spanning across many faiths, cultures and religions of the world. This is the principal database.
2. OMBIO: A database of life sketches/descriptions of the Sources with pictures where available (about 100 at present).
3. OMBIB: A bibliography of books for further reading (about 250 at present).

Records are added to the databases approximately every six months. The input to all the three databases are from corresponding databases prepared using CDS-ISIS software (of Unesco). The CD-ROM version has the Windows version of CDS-ISIS ported on to it. The usual Boolean, parenthesis, truncation and other operators can be used in the search.

In the OM database the search is essentially concept-based which can be qualified by other concept terms or name(s) of Source(s).

The PC version does not, at present, provide for retrieval of images and sound (speech and music). The PC and CD-ROM versions use a Pascal interface [2] to inter-link the databases, and to search in one or more databases concurrently. A user can select any term or terms from a displayed extract record, formulate a new search expression with them, and continue the search in the same database or in the other databases.

3. Indexing

The search and retrieval are based mainly on the indexes to the databases. An online thesaurus is in preparation. The fields of the databases and how they are indexed are mentioned below.

3.1 OM Database

The following fields are indexed:
Subject Heading (Whole heading and each substantive term separately)
Text of Extract (Each substantive term)
Source Name (Whole name)
Context (Each substantive term)
Notes (Each substantive term)
Verse number (for extracts from Bhagavad Gita)

3.2 OMBIO Database

The following fields are indexed:
Name of Source (Whole name)
Period of Source (Whole, e.g., 16th cent; 15th - 16th cent); specific dates of birth and death are not indexed.
Life Sketch (Each substantive term)

3.2 OMBIB Database

The usual Author, Title, Descriptor, Subject Heading, Abstract, etc., fields are indexed. Each substantive term of the Title and Abstract fields is indexed.

3.3 Language

Only English language materials (that is, the original is in English or is a translation into English) have been scanned in preparing the records (also in English). This raises some of the problems mentioned below.

3.4 Types of Queries to which OMIS Responds

Typical queries that OMIS responds to include:

- Texts of what St. John of the Cross and Soren Kierkegaard have said about "control of the senses" or "purity of heart".
- What does the Bhagavad Gita say about Karma? Interpretations of the relevant texts by S. Radhakrishnan and Anne Besant.
- The relation between "salvation" and "renunciation" as propounded by Thomas Merton, Meister Eckhart, Sri Ramakrishna Paramahamsa, Sri Sankara, and Sathya Sai Baba.
- Biographical sketch and picture of Mother Teresa of Calcutta.
- Reading materials on Sufism and Sufi Saints.

4. Vocabulary Control: Usual Requirements

Vocabulary control is necessary within a database and across databases, be it the name of a person, corporate body, or name of a subject/concept to facilitate search across databases using one and the same search expression. In regard to name of Source, the rendering in all the databases is according to some widely accepted cataloguing code, such as AACR2. Alternative names, pseudonyms, popular names, etc., are found in spiritual and religious texts also.

For name of subject/concept, again consistency is required not only in all the indexed fields within a database but also across databases, to facilitate hypertext links. The usual problems of spelling variation, grammatical variation, compounded words, synonyms, hyphenation, abbreviations, etc., are all encountered. These are not considered here. Only some additional issues relating to culture-sensitive terminology and LRs in developing the thesaurus to assist online searching in such databases are mentioned.

5. Lateral Relations

5.1 Levels of Lateral Relations

Lateral relations can exist or be perceived at different levels:

5.1.1 LR-3

When a user identifies several sources, such as websites, as possible sources likely to provide information to a query on hand, those sources are related with respect to the specific information need. This is inter-websites or inter-sources relation. This knowledge can be of help in further searches.

5.1.2 LR-2

When several databases are available, identifying those to be searched in relation to a specific information need depends on knowledge of the contents of each of the databases. When a user selects a set of databases for further search, he/she perceives a relation among them. This is inter-databases relation. This knowledge can be of help in subsequent searches in those databases.

5.1.3 LR-1

A search expression, for example, "Mysticism and Christian and Middle Ages", applied to a database may retrieve one or more texts or records. This establishes a relation among the records retrieved in relation to the particular query. This is inter-records relation.

5.1.4 LR-0

Consider the concepts represented by the terms Mysticism, Christian, Middle Ages, in the search expression mentioned in the previous section. The search may identify a similar relation among the concepts in the records retrieved and therefore deemed relevant to the query. This is inter-concepts relation. We shall be considering this type of LR in more detail.

5.2 Equivalent and Near-Equivalent Concepts

Most of the special concepts in the spiritual and religious domains are common to many faiths and cultures, for example, God, liberation, soul, prayer, after-life, bliss, etc. The terms used in the different faiths and cultures for the concept denoted by each of the above English terms, for instance, may not be exact equivalents or are not coterminous in what they signify or refer to. Other examples: Salvation, Moksha, Immortality, and Nirvana. This LR should preferably be distinguished from the usual RT relation. The vocabulary control tool used (e.g., thesaurus) may bring together these near-equivalents, using the symbol NE or ~ for Near-Equivalence. Example:

LIBERATION	ENLIGHTENMENT
NE ENLIGHTENMENT	NE IMMORTALITY
IMMORTALITY	LIBERATION
MOKSHA	MOKSHA
NIRVANA	NIRVANA
SALVATION	SALVATION

Similarly, with each of the other terms Immortality, Moksha, Nirvana, and Salvation.

Again, the terms Atman, Brahman, Self, and Soul are used as near-equivalents.

ATMAN	BRAHMAN
NE BRAHMAN	NE ATMAN
SELF	SELF
SOUL	SOUL
SELF	SOUL
NE ATMAN	NE ATMAN
BRAHMAN	BRAHMAN
SOUL	SELF

In texts, written in English or translated into English, one finds translated or transliterated terms, such as Karma, Atman, Brahman, Nirvana, Einsof, and Sefirot. In some texts a term considered to be equivalent by the translator may be used. In others the transliterated original term may also be given in parenthesis or in a note. In the OM database the English term and the transliterated term will automatically be indexed. A user may use such terms, that is, English or transliterated, in the search expression.

One and the same English term may be used in a translated text to denote slightly different concepts for which there may be different terms in the original work. For example, the term Mind used for Manas and Chitta, which may be given in parenthesis or in a note in the text. A reverse case is one in which the same term in the original (transliterated into English) may denote slightly different concepts for which different English terms may be used in the text.

For example, the transliterated term Buddhi is used for Reason, Intellect, and Intelligence; Prajna for Mind and Understanding (on the other hand, Intelligence may refer to Chetana). These variations may be explained in parentheses in the text or in a note to the text. In the OM database all the terms will be indexed automatically. A searcher can use any of the terms selected from the index and then select other terms from the text displayed and continue the search for further records, a facility provided by the Pascal hypertext link in the CD-ROM and PC versions. Again, in the thesaurus these terms are to be brought together as NE type RTs. In any case the user may need guidance to the different senses in which a term (whether an English term or a transliterated one) may be used in the text, within the same faith or culture and/or across them.

Equivalence Relation is usually a See cross-reference in subject heading lists or USE and UF direction in thesauri. For example:

AVES	AVES	BIRDS
See BIRDS	USE BIRDS	UF AVES

In culture-sensitive databases, the preferred term used for search depends on the culture / faith and/or linguistic background of the user. It would be expedient to provide for search using any of the equivalent terms, indicating in the vocabulary control tool (e.g., thesaurus) the nature of LR as EQ or = for Equivalence. Here are some examples:

MONASTIC LIFE
EQ SANNYASA

SANNYASA
EQ MONASTIC LIFE

GOD
EQ ABSOLUTE
SPIRIT
ALLAH
CREATOR
DIVINITY
EINSOF
INFINITE
ISWARA
PROVIDENCE
SUPREME
BEING

ALLAH
EQ ABSOLUTE
SPIRIT
CREATOR
DIVINITY
EINSOF
GOD
INFINITE
ISWARA
PROVIDENCE
SUPREME
BEING

EINSFOF
EQ ABSOLUTE
SPIRIT
ALLAH
CREATOR
DIVINITY
GOD
INFINITE
ISWARA
PROVIDENCE
SUPREME
BEING

Similarly, for each of the other terms.

HOLY WORD (The Bible)
EQ KALMA (Islam)
SABD
(Indian scriptures)
SAUT-E-SARMADI
(Sufi)
TAO
(China)
UDGIT
(Upanishads)

UDGIT (Upanishads)
EQ HOLY WORD
(The Bible)
KALMA
(Islam)
SABD
(Indian scriptures)
SAUT-E-SARMADI
(Sufi)
TAO
(China)

Similarly, for the other terms.

5.3 Attribute

Concepts representing attribute, characteristic, or property of an entity occur frequently. For example, Renunciation is an attribute of Monastic life, Yogi, Spiritual life, etc. This can be represented as follows:

MONASTIC LIFE
EQ SANNYASA
(Attribute)
RT RENUNCIATION

SELF-LESS ACTION
(Attribute)
RT RENUNCIATION

SPIRITUAL LIFE
(Attribute)
RT RENUNCIATION

YOGI
EQ SANNYASI
(Attribute)
RT RENUNCIATION

RENUNCIATION
(Attribute of)
RT MONASTIC LIFE
SANNYASA
SANNYASI
SELF-LESS
ACTION
SPIRITUAL
LIFE
YOGI

Another example:

GOD
(Attribute)
RT ABSOLUTE
UNDIFFERENTIATION
BOUNDLESS
CHANGELESS
ESSENCE OF
EXISTENTS
INCOMPREHENSIBLE
INEXPRESSIBLE
INFINITENESS
NON-DUALITY
OMNIPRESENCE
OMNISCIENCE
UNIFIED ONENESS
WITHOUT
BEGINNING

ABSOLUTE
UNDIFFERENTIATION
(Attribute of)
RT GOD

BOUNDLESS
(Attrihute of)
RT GOD

CHANGELESS
(Attribute of)
RT GOD

NON-DUALITY
(Attribute of)
RT GOD

etc.

5.4 Attained Through/Leads to

Method of or Approaches to achieving or attaining a goal or an object is a frequently occurring relation. Here are some examples:

RENUNCIATION
(Attainment through)
RT BHAKTI
DEVOTION
JNANA
SACRIFICE
SELFLESS ACTION
SELF-ANALYSIS
TURNING MIND INWARD
UNCEASING WORK
YOGA

BHAKTI
(Leads to)
RT RENUNCIATION

JNANA
(Leads to)
RT RENUNCIATION

SACRIFICE
(Leads to)
RT RENUNCIATION

etc.

ENLIGHTENMENT
(Attainment through)
RT RENUNCIATION

ETERNAL, THE
(Attainment through)
RT RENUNCIATION

IMMORTALITY
(Attainment through)
RT RENUNCIATION

LIBERATION
(Attainment through)
RT RENUNCIATION

RENUNCIATION
(Leads to)
RT ENLIGHTENMENT
ETERNAL, THE
IMMORTALITY
LIBERATION
MOKSHA
NIRVANA
PEACE
SALVATION

etc.

5.5 Comparison, Differentiation, Influence

Consider the following:

Comparison of Unselfish Action with Renunciation

The difference between Relinquishment and Renunciation

Influence of Meditation on Mental Stress

These relations can be represented respectively as follows:

RENUNCIATION
(Compared with)
RT UNSELFISH ACTION

UNSELFISH ACTION
(Compared with)
RT RENUNCIATION

RENUNCIATION
(Differentiated from)
RT RELINQUISHMENT

RELINQUISHMENT
(Differentiated from)
RT RENUNCIATION

MEDITATION
(Influence on)
RT MENTAL STRESS

MENTAL STRESS
(Influenced by)
RT MEDITATION

5.6 Model/Case Study

Discussion or discourse on a person as a Model of, say, true Renunciation or Yogi is another LR occurring in the extracts in the spiritual domain. This may be represented as follows:

RENUNCIATION
(Model)
RT Ramakrishna Paramahamsa
Ramana Maharshi

RAMAKRISHNA PARAMAHAMSA
(Model of)
RT RENUNCIATION
YOGI

YOGI
(Model)
RT Ramakrishna Paramahamsa
Ramana Maharshi

RAMANA MAHARSHI
(Model of)
RT RENUNCIATION
YOGI

5.7 Cause of

The concept of an entity having the capacity to Cause or Generate another entity also gives rise to a LR. For example:

EINSOF
SN The Infinite in Kabbalah, Jewish mysticism
(Causes)
RT SEFIROT

SEFIROT
(Caused by)
RT EINSOF

It is not clear whether the ten emanations (Sefirot), Keter, Tif'eret, Yesod, Shekkinah; Hokhmah, Hesed, Netsah; Binah, Geruvah, Hod; and Malkhut should be considered as hierarchical relation or as LR to Sefirot.

6. Work in Progress

Other types of LR occurring in spiritual and religious texts are being examined and will be compared with the thirty types identified more than two decades ago. How a faceted classification scheme, such as the Colon Classification, handles such relations is also being studied.

7. Remarks

Increasingly information access is becoming global. A wide range of information materials are available and a wide variety of users access, for example on the Internet, the globally accessible databases. In the type of databases in the spiritual domain discussed here, terms occurring in the texts and those used for searching are culture-sensitive. Therefore, in a database that uses only English language source materials, use of English language terms for indexing with, say, bias to a particular culture or faith is inexpedient. As discussed in this presentation, providing for search using terms of different cultures (e.g., in transliterated form) is necessary.

The reasons for an English text (either original or translated) for using transliterated terms, such as Dharma, Karma, Nirvana, Einsof, Shekkinah, etc., and giving the meaning or a term deemed to be equivalent or near-equivalent in parenthesis or in a note could be that the English term is less widely known or used, or the English term is only a near-equivalent to the original term, or an English equivalent does not exist or not known to the author. Similar reasons apply to an information seeker to use a transliterated term in the search. The source text may adopt the reverse of the above giving the transliterated term in parenthesis or in a note. In the OM database all the text terms including those occurring in parenthesis and in the notes are indexed. The need to provide the user approaches from different forms of a term has already been mentioned above.

We would like to know of the experiences and methods adopted by the participants who have developed similar multicultural, global information systems.

References

Rajashekar, T. B., Ravi Srinivas & Neelameghan, A. (1998). Designing a multimedia information service for the Internet and CD-ROM. *Information Studies, 4(3).*

Srilatha, G. & Neelameghan, A. (1995). A Microisis Pascal interface for concurrent multiple databases search and retrieval. *Information Studies, 1(2).* 114-129

Professor Hudon is on the faculty of the École de bibliothéconomie et des sciences de l'information (EBSI), Université de Montréal.

Compatibility and identity are not synonyms : Conceptual structures in multilingual thesauri

Michèle Hudon

Put side by side several monolingual, independently developed thesauri describing the same field in different languages, and one of the first things you will notice is that their respective structures are not identical even if they are built around the same concepts. A possible explanation for this is that their respective developers took some liberty in applying the guidelines (I think such a hypothesis would be easy to verify!). But there is also something else: thesaurus developers with different linguistic backgrounds and coming from different cultural environments do not see and organize the world in the same way. Anthropologists, translators, linguists ... and any tourist will attest to the fact that relationships in knowledge organization, as in daily life, are affected and constrained by linguistic and cultural contexts.

In multilingual thesauri, descriptors considered as linguistic equivalents do not necessarily refer to the exact same concept or cover the exact same area in the conceptual space. It stands to reason, then, that relations between concepts within languages will also vary. So why is it that in so many multilingual thesauri today, and not the most obscure ones, identity of hierarchical and associative structures still appears as such a desirable characteristic? These instruments would have you believe that multilingual thesauri are language- and culture-neutral.

A quick look into the past may help us understand why things are as they are now.

The first experiences of interlingual information transfer required a huge effort of harmonization and standardization, and controlled vocabularies offered a workable solution to many problems associated with the process. The first multilingual thesauri were developed rapidly in the seventies, when the progress of technology made very real the prospect of a global information system; there was no question that such a system would be multilingual.

In multilingual thesauri, much emphasis was put on compatibility of structure: strong compatibility resulted from full correspondence of concepts and relations, while weak compatibility resulted from correspondence between concepts but not between conceptual relations.

Thesaurus developers were advised that "international comparability and practical applicability [were] far more important than absolute conceptual correctness" (Beling et al., 1974) in multilingual thesauri.

Thesaurus workers relied on guidelines that stated that "as a general rule, any hierarchy which the users of one language regard as logically acceptable should be equally valid when its terms have been translated into another language" (International Organization for Standardization, 1985). They moved quickly over warnings like this one: "Before an associative relationship which has been recognized in one language is transferred to another, it should be examined to determine how far it continues to be valid; if it appears to apply to only one group of language users, it should generally be excluded" (International Organization for Standardization, 1985). They agreed with their mentors who believed that "the problems of multilingual thesaurus construction [were] no worse, in kind, than those of monolingual thesaurus construction" (Aitchison & Gilchrist, 1987, p. 108), and assumed that the most difficult aspect of the work would be that of human organization.

The software that was developed to facilitate the task of building thesauri took the notions of source and target languages very far, in fact providing for the creation of a monolingual structure, and then generating different linguistic versions using a basic file of equivalents, in accordance with a model proposed by Rolling (1979), and experimented early on in the European Community.

Controlled vocabularies were designed for specialists who generally accepted to work within the constraints imposed by the system. One such constraint, for minority language searchers, was the somewhat artificial character of the searching language at their disposal. Interlingual communication had been achieved, but at the expense of intercultural communication.

There were irritants though, voices coming from the social science and humanities communities (where concepts are very much culture dependent and special languages are closer to natural languages than in the scientific domains), as well as voices from non English-speaking communities and from Eastern cultures. A project of adapting the *International Thesaurus of Cultural Development* provided a platform for those who advocated more flexibility in conceptual structures and true cultural representativeness; in the end, however, existing practices were not much affected.

Today' discourse emphasizes "ross-language communication" and "ultilingual access to multilingual information" rather than the "anguage barrier".The objectives of the Multilingual Information Access (MLIA) project are representative of many others in the same area. They are:

1. To allow individuals to use the language that they feel most at ease with so they can formulate queries as simply and intuitively as possible;
2. To provide interpretation support to access information within documents written in a foreign language (EU-NSF Collaborative working group 1998).

Natural language processing (NLP) in a multilingual environment is at the core of major research efforts at this time. Not surprisingly, polysemy seems to become a problem more rapidly in multilingual than in monolingual contexts, and a reliable method for sense disambiguation must be found. The controlled vocabulary approach, because it provides a context of sort, is still considered appropriate for such disambiguation, and it is now coupled with techniques based on corpus statistics.

There is a future for the multilingual thesaurus, but not, I suggest, for the thesaurus in the exact form in which it has existed for the past 30 years: the 'do nothing scenario' is not a valid option anymore, given the size and the reach of the global network, the levels and new characteristics of the users, the necessity to provide for other than hierarchical ways of organizing knowledge.

Much creativity has been applied to finding solutions to problems linked to relations between concepts and their verbal representations (the problems of interlingual equivalence). But if we want the multilingual thesaurus to remain useful to a large base of potential users, if we want it to serve as semantic map in multilingual NLP systems, we must re-visit the problems and past decisions regarding its conceptual structure.

Research on multilingual thesauri structuring and applications must move further into the following two directions:

1. *A search for truly common conceptual structures.* Thesauri have traditionally taken a light approach to relationships, using only fairly general and comprehensive relations between concepts. This may have been sufficient in small thesauri, but in the highly specific thesauri of today (whether monolingual or multilingual), there is a need for a refinement in the definition of relationships which will inevitably lead to structural differences across languages. There is probably no limit to the distinctions that could be made within sets of associative relationships, and the challenge will be to identify the types of useful relations that are perceived the same way and have the

same value for sense disambiguation in various linguistic contexts. Such a refinement in the definition of relationships will lead to a more complex thesaural structure; let's not forget, however, that our thesauri will increasingly be used by machines for query analysis, interpretation and expansion, and those machines need very clear and complete semantic maps. It is significant that a major difference between the EuroWordNet database and its English-only counterpart WordNet is a more refined network of relations among concepts. It is also interesting to note that EuroWordNet differentiates relationships that are subsumed into one or two in thesauri.

2. *Taking advantage of the technology.* Having found a common structure, we could continue to produce identical structures in all linguistic versions of a multilingual thesaurus. But there might not be a need to do so.

Why not take advantage of available technology and software that allow us to maintain separate structures connected by some form of language neutral switching mechanism? One recognizes here again the EuroWord Net approach (the equivalent in fact of what some thesaurus developers have been doing manually for years because no software would tolerate gaps in hierarchies or admit other structural variations within linguistic versions). In EuroWordNet, each monolingual file reflects semantic relations as a language-internal system, maintaining cultural and linguistic differences while still providing for cross-language exchanges (Vossen et al., 1997). In computational linguistics, more flexible processing models are being tested, permitting more than one view of the world to coexist in the final product, and giving this product a more democratic touch.

The technology allows us, finally, to get away from the belief that compatibility in multilingual thesauri must of necessity be equated with identity of structure, leaving us free to concentrate on cultural representativeness, usability and user-friendliness, and maybe, just maybe, giving all users an equal chance of finding interesting material when they dare 'express their query intuitively'.

It is hoped that the next edition of the guidelines for the development of multilingual thesauri will reflect these new circumstances.

References

Aitchison, J. & Gilchrist, A. (1987). *Thesaurus construction: a practical manual*. London: ASLIB.

Beling, G., Schuck, H. J. & Wersig. G. (1974). Procedural guide for the translation of foreign language thesauri into German. In *International scientific symposium on multilingual thesauri, 8-10 October 1973: Proceedings.* Berlin: Leitstelle Politische Dokumentation. 55-114.

EU-NSF Collaborative working group. (1998). *Multilingual information access (MLIA): Executive summary of white paper.* http://www.cs.columbia.edu/~klavans/Activities/MLIA/98-MLIA-WhitePaper-ExecutiveSummary.html

Fluhr, C. (1996). Multilingual information retrieval. In *Survey of the state of the art in human language technology.* http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html

International Organization for Standardization. (1985). *Guidelines for the establishment and development of multilingual thesauri (ISO 5964-1985).* Geneva: ISO.

Maniez, J. (1997) Fusion de banques de données documentaires et compatibilité des langages d'indexation. *Documentaliste – Sciences de l'information, 34(4-5).* 212-222.

Mirbel, I. (1997). Semantic integration of conceptual schema. *Data and Knowledge Engineering, 21(2).* 183-195.

Oard, D. & Dorr, B. J. (1996). *A survey of multilingual text retrieval* (UMIACS-TR-96-10; CS-TR-3615). http://www.ee.umd.edu/medlab/filter/filter_project.html

Peters, C. & Picchi, E. (1997). Using linguistic tools and resources in cross-language retrieval. In *Third Delos workshop on cross-language information retrieval, Zurich 5-7 March 1997.* http://www.ee.umd.edu/medlab/filter/sss/papers

Raybeck, D. & Herrmann, D. (1990). A cross-cultural examination of semantic relations. *Journal of Cross-Cultural Psychology, 21(4).* 452-473.

Rolling, L. (1979). Computer management of multilingual thesauri. In *Ordering systems for global information networks: Third international conference on classification research, Bombay, 6-11 January 1975.* Bangalore, India: FID. 382-388,

Sedelow, W. A. & Sedelow, S. Y. (1994). Multicultural/Multilingual electronically mediated communication. *Social Science Computer Review, 12(2).* 242-249.

Soergel, D. (1997). *Multilingual thesauri in cross-language text and speech retrieval.* In *AAAI Spring Symposium on Cross-language text and speech retrieval, March 24-26 1997: Electronic working notes.* http://www.ee.umd.edu/medlab/filter/sss/papers

Vossen, P. et al. (1997). The multilingual design of the EuroWordNet database. In *IJCAI-97 Workshop on ontologies and multilingual NLP, Nagoya, Japan, Saturday August 23, 1997.* http://crl.nmsu.edu/Events/IJCAI

Professor Fluhr is on the faculty of the Institut National de Sciences et Techniques Nucléaires (INSTN) and serves as an advisor to the Direction de l Information Scientifique et Technique, Commissariat à l Energie Atomique (DIST-CEA).

Lexical Knowledge and General Public Online Search

Christian Fluhr

1. Internet and the End-User, a Moving World

The use of the Internet, giving access to information to a much larger public, has the following consequences:

Controlled vocabulary is replaced by natural language uncontrolled vocabulary

Keywords and abstracts are generally replaced by full text

Boolean interrogation is replaced by natural language interrogation

Monolingual interrogation is insufficient in more and more cases, resulting in a need for crosslingual retrieval

Very few Internet search engines use more than character string search, sometimes simulating stemming by automatic truncation. Indeed, the few that incorporate lexical knowledge use implicit relations between the words in the documents to support relevance feedback (e.g., Live topics within Altavista).

In fact, linguistic processing is required to fully exploit lexical knowledge with maximum precision, but linguistic indexing cannot be done on volumes like the full Internet. But there is also a need for intranet application where the volume of data is compatible with the use of sophisticated linguistic indexing and retrieval.

2. Problem of the Construction of Uncontrolled Vocabulary

For general language some large multinational projects like WordNet and its European multilingual counterpart EuroWordNet have been launched.

For domain vocabulary, a large amount of work must be done :

- *To extend and modify existing ontologies like thesauri and dictionaries* (monolingual and bilingual). In fact, thesauri cannot be used without modifications. Words are not normalized as they are in natural language processing (e.g., upper case characters appear without diacritics; sometimes words are in the plural form; sometimes the word is qualified to disambiguate it). It is the same for relationships: a BT relationship between a compound and its head is automatically given by syntactic parsing and is not useful to incorporate into the system's lexical knowledge; synonymy between different forms of an acronym (e.g., CEA, C.E.A.), between different forms of a compound (e.g., payload, pay-load), or between words having the same root can be obtained automatically by linguistic processing.
- *To discover and use in searching more sophisticated relations than the ones now found in thesauri* such as kind of, part of, agent of (an action), object of (an action), instrument of (an action), etc. ...

3. Problems in Using Uncontrolled Vocabularies

The main problem in using uncontrolled vocabulary is the problem of ambiguity. This problem can be addressed by :

Morphosyntactic analysis on both texts and queries

Semantic analysis (only if it can be applied on general vocabulary)

Use of implicit lexical semantic knowledge in the database: the database can be used as a semantic filter to improve relevance of the query answer

To solve ambiguity, context is needed. Thus, the description of the search topic must be of sufficient length to facilitate disambiguation. This is important, and users must exhibit the completely opposite behavior from that which they used with Boolean queries. For Boolean queries the query must include a minimum of words with the AND operator to ensure having a nonempty answer. In natural language queries, the longer the query is, the larger the context for disambiguating ambiguous words. This means that precise queries to access precise localized information can give better results.

4. Understandability of the System Behavior

Users want the system to explain why it proposes that documents are relevant.

Most of the purely statistical systems that use implicit lexical semantic knowledge from the database are

unable to explain to a human why a document is relevant even if this kind of system can give very good results (see Latent Semantic Indexing systems in TREC).

The only way is to use explicit relations built manually or even automatically built from the database.

5. Multilingual Information

The globalization of the economy brings more and more need for access to multilingual information. The dream of a world where every economic actor speaks only English is no longer pursued, even in the United States. The White House has imposed translingual information management as a main theme of research cooperation between the U.S. and the European Union.

Three ways of solving the problem of cross lingual interrogation are being explored:

- Use of the statistical approach – based on the existence in the database of translated documents or documents about the same events – to discover implicit relations between words in the corpus.
- Use of machine translation (MT) systems, but MT systems are weak with respect to semantic disambiguation; this can result in low recall.
- Use of bilingual reformulation and semantic disambiguation by the database; this is the more promising approach, but needs high quality lexical knowledge.

6. Conclusion

Rapid changes in tools and habits are sometimes hard to manage. The effectiveness of systems for end users depends strongly on the effort undertaken by domain specialists on the system's lexical knowledge. We face a challenge like the one we faced when we constructed thesauri, but now the step is higher.

References

Fluhr, C., Schmit, D., Elkateb, F., & Gurtner, K. (1997). Multilingual database and crosslingual interrogation in a real Internet application. Workshop "Cross-language Text and Speech Retrieval" in "AAAI 1997 Spring Symposium Series", 24-26 March 1997, Stanford University, California.

Debili, F., Fluhr, C., & Radasoa, P. (1988) About reformulation in full text IRS. Conference RIAO 88, MIT Cambridge, March 1988; a modified text was published in *Information Processing and Management*, 25/ 6 (1989), 647-657.

Grefenstette, G. and al. (1998). *Cross-language information retrieval*. Boston: Kluwer Academic Publishers.

Ms. Mitchell is editor of the Dewey Decimal Classification (DDC).

Flexible Structures in the Dewey Decimal Classification

Joan S. Mitchell

Our panel has been asked to address the limits and potential of hierarchical structures; the constraints on relationships posed by language, culture, and discipline; and the role of relationships in the online environment. I will address all three of these areas in a discussion of how flexible structures could be used to transform a general library classification scheme such as the Dewey Decimal Classification (DDC) into a general knowledge organization tool for the worldwide electronic information environment.

In the extended version of his address to ISKO 4, Fran Miksa (1998, p. 89) calls for making the DDC into a more malleable system than it is at present: "We will have to see the entire system as a vast array of moveable or interchangeable facets of categories ...." How do we achieve such flexibility in a seemingly discipline-bound hierarchical structure with Western culture and language biases? I will describe some of the ways this challenge may be addressed through flexible structures that co-exist with the general scheme.

What do I mean by flexible structures? A flexible structure is an alternative view that is derived from or linked to a general organization scheme to address an information need not easily accommodated through the existing structure. Some flexible structures already exist within the scheme but have not been exploited due to the limits of current retrieval mechanisms. For example, there is untapped potential within the notation and in the polyhierarchical links resident in the DDC. Last summer, I participated in a panel discussion at Lund University in which one of the speakers observed that current classification schemes do not support hypertextual browsing (Lundberg, 1997). They do; we just have not exploited this feature. For example, in the number for the topic "respiration in bats," two different hierarchies are linked together and are available for searching – the hierarchy for respiration in physiology, and the hierarchy for bats in mammals:

Respiration in bats 573.2194
573.2 Respiration
1 Facet indicator for specific animals (from 571.1 Animals)
94 The number that follows 59 in 599.4 Bats

In addition, the use of uniform notation for bats disambiguates bats in the sense of mammals from bats in the sense of "baseball equipment" in other hierarchies in the DDC:

Bats	599.4
conservation technology	639.9794
paleozoology	569.4
resource economics	333.9594

not

Bats (Baseball)	796.35726
manufacturing technology	688.76357

Several years ago, Liu (1993) demonstrated the feasibility of "decomposing" Dewey numbers in the 700s into their component parts. Later in this conference, Steve Pollitt (1998) will describe his research on view-based searching using Dewey facets in an online catalog.

The full and abridged editions of the DDC have always included numerous optional arrangements to address the special needs of users due to cultural differences or differences in the quantity or nature of the literature. Options provide alternatives to the standard structure in terms of jurisdictional emphasis; racial, ethnic, national group emphasis; language emphasis; topical emphasis; or emphasis by some other special characteristic (Mitchell, 1995). The various translations of the DDC often include adaptations and expansions to address various cultural needs. These alternative views are useful, but again, they usually address the needs of the general user in another cultural setting. What about the needs of specific discourse communities in a general scheme? Here, the introduction of a virtual flexible structure through the overlay of a different vocabulary and structure is important. It may be a formal structure, such as another thesaurus, or a user-defined structure, such as the "Knowledge Class" structure proposed by Lin and Chan (1997), or even the structure of a search engine such as Yahoo!.

I will briefly describe the record we have developed to accommodate the mapping of vocabulary from another structure. Later, Hope Olson (Olson & Ward, 1998) will describe a research project in which another thesaurus, *A Women's Thesaurus*, is linked to the DDC to provide an extended vocabulary and an alternative view, or flexible structure, for the discourse community of women's studies.

To support flexible structures within the DDC, we have developed an authority control module for entries in the Relative Index and for linked entries from other thesauri. Each record accommodates the following information:

- index entry
- links to schedule, table, and Manual records
- editorial note
- scope note
- source data found (similar to 670 field in authority format)
- source data not found (similar to 675 field in authority format)
- confidence level
- 70X-75X index term fields (from the MARC classification format)
- cross references (with the nature of the reference labeled: BT, NT, RT, UF, USE)

This record accommodates the mapping of equivalent concepts from other thesauri to Relative Index terms, and also accommodates the mapping of concepts from other thesauri directly to the DDC.

What are the open research questions in the mapping of one structure to another structure? The most obvious is the definition of the relationship. We are experimenting with a simple set of three relationships in the "confidence level" field for selective mapping of Library of Congress subject headings to the DDC (Mitchell, 1996):

(1) This heading points to this number exclusively
(2) This heading maps to this number and others
(3) Other

Several years ago, Iyer and Giguere (1995) suggested seven relationships for the linking of the American Mathematical Society Mathematics Subject Classification to the DDC:

(1) Exact matches
(2) Specific to general
(3) General to specific
(4) Many to one
(5) Cyclic mapping strategies
(6) No matches
(7) Specific and broad class mapping

Meo-Evoli, Negrini, and Farnesi (1998) will also address definitions of relationships between knowledge structures later in the conference. There is much work to be done on coding the nature of relationships, including investigating how the purpose for which the relationship is to be used affects the definition of the link.

I will close by observing that our present knowledge organization systems have explicit and implicit features

for supporting browsing and retrieval in the online environment. Hierarchy can always play a useful role when one is not sure of the name of a concept, or the concept has an ambiguous name or is known by several names. Our existing knowledge organization structures need to be mined for the additional information resident across the hierarchies, and to be extended by the overlay of formal or informal knowledge structures to make them useful tools in the online environment.

References

Iyer, H. & Giguere, M. (1995). Towards designing an expert system to map mathematics classificatory structures. *Knowledge Organization* 25(3/4): 141-147.

Lin, X. & Chan, L. M. (1997). Knowledge class: A dynamic structure for subject access on the web. In E.N. Efthimiadis (Ed.). *Proceedings of the 8th ASIS SIG/CR Classification Research Workshop, November 2, 1997, Washington, D.C.* Silver Spring, Md.: ASIS. 31-40. Also published in E.N. Efthimiadis (Ed.). (1998). *Advances in Classification Research, vol. 8: Proceedings of the 8th ASIS SIG/CR Classification Research Workshop.* Medford, N.J.: Information Today. 33-42.

Liu, S. (1993). The automatic decomposition of DDC synthesized numbers. Ph.D. diss., University of California, Los Angeles.

Lundberg, S. (1997). Untitled paper read at the seminar, Electronic Challenge, 28 August, at Lund University, Sweden.

Meo-Evoli, L., Negrini, G. & Farnesi, T. (1998). ICC and ICS: comparison and relations between two systems based on different principles. In W. Mustafa el-Hadi, J. Maniez & A. S. Pollitt (Eds.). *Structure and Relations in Knowledge Organization: Proceedings of the 5th International ISKO Conference.* Würzburg: Ergon Verlag. 229-237.

Miksa, F.L. (1998). The DDC, the universe of knowledge, and the post-modern library. Albany, N.Y.: OCLC Forest Press.

Mitchell, J.S. (1996). The Dewey Decimal Classification at 120: Edition 21 and beyond. In R. Green (Ed.) *Knowledge Organization and Change: Proceedings of the 4th International ISKO Conference.* Frankfurt: Indeks Verlag. 378-385.

Mitchell, J.S. (1995). Options in the Dewey Decimal Classification system: The current perspective. *Cataloging & Classification Quarterly* 19(3/4): 89-103. Also published in A. R. Thomas (Ed.). *Classification: Options and opportunities.* New York: Haworth. 89-103.

Olson, H.A. & Ward, D.B. (1998). Charting a journey across knowledge domains: Feminism in the Dewey Decimal Classification. In W. Mustafa el-Hadi, J. Maniez & A. S. Pollitt (Eds.). *Structure and Relations in Knowledge Organization: Proceedings of the 5th International ISKO Conference.* Würzburg: Ergon Verlag. 238-244.

Pollitt, A.S. (1998). The application of Dewey Decimal Classification in a view-based searching OPAC. In W. Mustafa el-Hadi, J. Maniez & A. S. Pollitt (Eds.). *Structure and Relations in Knowledge Organization: Proceedings of the 5th International ISKO Conference.* Würzburg: Ergon Verlag. 176-183.

Dr. Bean is at the Lister Hill National Center for Biomedical Communications, National Library of Medicine (NLM), working with the Unified Medical Language System (UMLS).

The Semantics of Hierarchy

Carol Bean

My remarks today derive from my experiences in medical informatics, but I believe the principles I address will generalize to other subject domains. Three basic assumptions underlie my own interests and research on semantic relationships; the first regards the importance of relationships; however, I don't think I need to belabor that point to this audience.

The second assumption asserts that controlled medical vocabularies are domain knowledge bases. A source of extensive structured domain knowledge is necessary for a variety of tasks. A domain model satisfies this need by defining the entities and relationships in some world. Characteristics of the best domain models include extensive breadth of coverage, relationships explicitly encoded as rules, and its entities are atomic concepts (or where complex, the internal relationships are explicitly defined). Domain models as we know them typically represent but a single perspective on a particular (single implied) domain. Controlled vocabularies are (underspecified) knowledge bases that provide one or more perspectives on a given subject domain from a particular point of view, in other words, a domain model. Vocabulary content is determined by the subject domain, and the organization of that content reflects a particular perspective on that domain, i.e., context. (The exact subset

of domain knowledge represented is also determined to some degree by the perspective.)

Knowledge structures are far more than the sum of their concepts. Their optimal construction and use in operation, as well as their integration across disparate systems and the sharing of knowledge therein, requires an explicit understanding of the organizational principles underlying their structure. Knowledge sources themselves contain at least some of the keys to integrating them one with another via their syndetic structure. My third assertion then is that much of the knowledge in a controlled vocabulary is contained within its syndetic, or relational, structure. The syndetic structure of a controlled vocabulary may then be seen as an organized expression of the relationships (equivalence, hierarchical, and associative) among its concepts, and used to discover implicit and to characterize explicit relationships. Precise specification of the myriad vocabulary structures in a domain will provide a contextual dimensionality for the knowledge contained in each that is sufficient to support integration across multiple views of a single domain, and indeed, across multiple domains.

In a given information system, the exact meaning of a concept is determined by the context in which it occurs; the relationships a concept has with other concepts in the system will define its context and thus its meaning. Contextual information in knowledge structures is most often conveyed via hierarchy. What principles the hierarchy and its subunits are organized around may be seen to reflect the predominant organizing principles of the domain itself.

Hierarchy has long been the dominant structuring mechanism in knowledge organization. There have been numerous efforts to inventory hierarchical relationships. While the resulting lists vary somewhat, most investigators would agree on the primacy and predominance of two hierarchical relationships. The most common, and perhaps quintessential, hierarchical relationship is IS-A, which describes the relationship between a class and a subclass or a type and its instantiation. The other primary hierarchical relationship is PART-OF, which most typically describes aggregation or composition.

Vocabulary and knowledge-base developers do not always distinguish between hyponymy (IS-A) and meronymy (PART-OF) in their hierarchies, often mixing them both among and within individual trees. Such "mixed" hierarchies prevail, or even predominate, in medical vocabularies and classification schemes (and from what I've seen, this situation obtains elsewhere as well). There are both advantages and disadvantages to mixed hierarchies. On the plus side, they provide a valid perspective on a subject domain and how the experts see it. However, there exist a multitude of risks arising from our ignorance of these structures and how they work.

One reason this is important is because of our current interest in computational use of existing knowledge structures; for example, ontologies form the core of knowledge-based information retrieval and natural language processing. Operations on knowledge structures depend on the characteristics of those structures. In general, we need to know what the relationships in a knowledge structure are; that is, what the specific individual relationships obtaining among various concepts are as well as the patterns within the knowledge structure as a whole. Still, this is not enough; we don' t know much about the relationships themselves, much less their behavior in different knowledge structures. What are the characteristics and properties of different types of relationships? How do relationships "work?" What is their functional or operating logic, and how does it vary? What relationships among relationships? How do characteristics of relationships affect their arguments or the entities they link? We also need to understand the characteristics of hierarchy as a knowledge structure, and its strengths and weaknesses. To sum: it is critical to understand how the characteristics of hierarchical relationships impact cognitive and computational operations.

Operations on hierarchies depend on several assumptions about the relationships they are structured around, and on two in particular. These are "directionality" expressed as some form of superordination and inheritance based on transitivity. Because hierarchy implies some sort of precedence or governance of one participant in the relationship over the other, each relationship asserted in a hierarchy can be seen to have an inherent and specific direction, which also defines the direction of inheritance. Reflecting these principles, class members typically display the characteristic attributes of the classes to which they belong; likewise, we expect subclasses to resemble their superclasses by virtue of attribute inheritance. The principles of transitivity and inheritance are the essential hallmarks of hierarchical knowledge structures, and are used extensively in operations on them. These properties make hierarchical taxonomy both a cognitively satisfying and a computationally powerful structure for organizing knowledge. The economy and efficiency they permit have made it the standard structure for knowledge representation.

Unfortunately, our computational and cognitive reliance on the principles of hierarchy in knowledge structures may be on shaky theoretical ground. It is clear that these principles operate differently among different relationship types. These properties reliably apply only to certain types of hierarchical relationships. For example, attributes of the whole may well obtain for its parts based on division (e.g., piece of pie), but not for its component parts (e.g., pie crust vs. pie filling). Further it is clear that many of the so-called Parent-Child relationships in knowledge structures are not prototypically hierarchical after all, which suggests both the potential fallacy of the assumptions as well as the necessary limitations surrounding their application. That these basic assumptions of hierarchy would be affected by mixed-relationship hierarchies and by non-hierarchical structures may be obvious, but precisely how, and perhaps more important, how they might be exploited is not, and remains to be discovered.

It is necessary to make explicit all interconcept links in a controlled vocabulary, even (especially!) the hierarchical ones, if we are to be able to exploit their inherent syndetic structure. After identifying what relationship types actually do exist in hierarchies and other knowledge structures, we may then determine their operating principles, such as their underlying logic. An increased awareness and understanding of such relationships and relational logic will inform our reliance on the assumptions underlying the basic principles of organization in knowledge structures and enable us to truly begin to approach knowledge-based information systems.

Dr. Williamson is Professor emerita, Faculty of Information Studies, University of Toronto.

Concluding Remarks

Nancy Williamson

At the end of this excellent conference it is time to reflect on what has been accomplished. Over four days, the participants have listened to 53 papers on various aspects of Structures and Relations in Knowledge Organization, which have been presented to them in 13 sessions. In support of the presenters – they have participated eagerly and enthusiastically in discussions both inside and outside of the conference room. The coverage of topics has been broad and complex, and in the time allotted it is not possible to cover all facets of the conference in detail or to do justice to all that has transpired. However, it is important to bring the conference to a close with at least a brief overview of our actions and to ask that fundamental question, "Where are we, and where should research in knowledge organization go from here?" In that respect there are a few observations that can be made.

First of all, there are a number of general trends that can be observed by scanning the topics from the five ISKO conferences held since 1990. Prior to the Lille Conference, I had been looking back in time, seeking some directions and inspiration for a theme that might be suitable for ISKO 6 in the year 2000. In doing so, a list of the categories from the previous four conferences was drawn up in order to determine whether or not there was any pattern in ISKO's accomplishments so far, that might suggest issues to be addressed in the future. The first discovery of note was that the number of categories of presentations have increased markedly in the period from 1990 at Darmstadt up to Washington in 1996. At Lille, there have been 11 categories in 13 sessions and a clear indication of where the emphases and the interests in issues lie. Three categories have stood out as requiring two sessions each – Cognitive Approaches, Linguistic Aspects and Design of Information Systems. Secondly, all of the conferences have reflected their themes to some extent, but some more faithfully than others. For example, at Madras in 1992 there was a fairly precise focus on the theme Cognitive Paradigms in Knowledge Organization. However, most of the conferences accommodated their themes while also providing variety in coverage. Overall there is considerable evidence of increasing diversity and breadth of coverage as one conference followed another. Traditional classification schemes still play a significant role in discussions, but in new and innovative ways. It now appears that classification and classificatory principles are understood as being fundamental to all kinds of information systems – a statement that probably could not have been made ten years ago. Moreover, the diversity of coverage has resulted in new contacts and new links with colleagues in other disciplines. Such diversity is a good omen and if followed through bodes well for the future of knowledge organization and ISKO.

As we observe the content of specific topics more closely, some other important trends are apparent. Theory has always played an important role in the ISKO Conferences. Nevertheless at Lille, there appears to have been a high degree of emphasis on theory, not only with respect to theory *per se* but also in the foun-

dations set out in the more application-oriented presentations. In this regard, two examples from the early sessions in the conference come to mind – Epistemology and Cognitive Approaches. It is to be expected that technology will play a more and more prominent role in research, as evidenced in the sessions on Computational Models and Automatic Domain Analysis. Interests in knowledge organization are gradually becoming more sophisticated. The most striking evidence of this is the very visible evidence of such topics as Linguistics and Cognitive approaches. Linguistics first appeared as a category at Copenhagen in 1994, while Cognitive approaches was first given prominence at Madras in 1992, but these topics really have come into their own in Lille. Traditional classification systems were there with emphasis on change and the problems of interdisciplinarity. This latter term first appeared in the categories at Washington in 1996, while more new ground was broken at this conference with the inclusion of papers on Visualization and Imaging. These new topics appear to be an important breakthrough and signify emerging areas for research. Other topics of interest and importance to the future are Ontologies, Conceptualization and Modeling. True to the theme of ISKO 5, Structures and Relations were topics that were well represented and debated in virtually every presentation made at the conference.

Finally, in the overview one can see that there is definite evidence of diversity, breadth and interdisciplinarity, but there is also some sense of *déjà vu*. Thus it is important to return to the initial question: "Where does knowledge organization research go from here?" The answer strikes a warning note. To have a lasting effect the results of research have to be cumulative and generalizable. The general impression, given past history and current discussions, is that we have not, at this point in time, achieved that goal.

We still need to gather that cohesive body of research which might contribute to a theory of knowledge organization. We are very concerned with individual problems without recognizing that they are symptoms of problems in knowledge organization in general. We need to broaden our horizons more to encompass knowledge organization in all disciplines. What characteristics of organization do those disciplines have in common? What can we learn as the basis for establishing a general theory? We may be on the way, but we are certainly not yet there.