

LAYERS OF MEANING: DISENTANGLING SUBJECT ACCESS INTEROPERABILITY

Joseph T. Tennis
Information School
University of Washington
Mary Gates Hall
Box 352840
Seattle, WA 98195-2840
jtennis@u.washington.edu

ABSTRACT

In order to facilitate subject access interoperability a mechanism must be built that allows the different controlled vocabularies to communicate meaning, relationships, and levels of extension and intension so that different user groups using different controlled vocabularies could access collections across the network. Switching languages, the tools of controlled vocabulary compatibility, consist of a single layer that does not allow for a flexible control of the semantic levels of meaning, relationships, and extension or intension. This paper proposes a multilayered conceptual framework wherein the levels of meaning, relationships and extension and intension are each controlled as individual parameters, rather than in a single switching language.

1. THE PROBLEM

In general, controlled vocabularies are built or adapted for a collection, in order to grant subject access to them. More and more controlled vocabularies are being used on the web (Koch et al., 1997) for the same purpose. To compound the problem standardized or universal schemes are not popular choices for specialized collections and so new schemes are built from the ground up. However, they are all different vocabularies built for different collections and different users. In an ideal situation, these different controlled vocabularies could work together, or interoperate, to provide subject access to resources beyond their own collections. This ideal situation is called *subject access interoperability*. In a formal sense, subject access interoperability is the state whereby different controlled vocabularies provide subject access to collections in a networked environment, beyond their own. Currently, this ideal does not exist. Thus the question surfaces of how can the state of subject access interoperability can be achieved.

2. PROPOSED SOLUTION

In order to facilitate subject access interoperability a mechanism must be built that allows the different controlled vocabularies to communicate meaning, relationships, and levels of extension and intension so that different user groups using different controlled vocabularies could access multiple collections in a networked environment. However, this is not a simple solution, nor a simple mechanism to build. Controlled vocabulary compatibility and conversion have been attempted from the advent of discipline specific thesauri (Dahlberg, 1996b; Dahlberg, 1996c), and work is still being done on issues of compatibility (Doerr, 2001). The past and present work on compatibility research strongly influences subject access interoperability, but varies from it through the control of variables such as meaning, relationships, and levels of extension and intension. Compatibility methods use a single layer to create a translation between participant

schemes, limiting the variables of control to that one layer. Thus that single layer must control meaning, relationships, extension, and intension variables. The conceptual framework of the subject access interoperability mechanism is multilayered – separating into many layers the components that control meaning, relationships, extension and intension. Subject access interoperability must offer a flexible control in this networked environment of different discourse communities. It will do so with layers.

3. PURPOSE OF THE PAPER

This purpose of this paper is to begin to explore the conceptual framework of a subject access interoperability mechanism. This requires a look into past work on controlled vocabulary compatibility and conversion. Each work reviewed below is fixed within a shallow taxonomy of compatibility or conversion methods. This taxonomy highlights how layers play a varied but vital role in the control of meaning, relationships, extension, and intension of each compatibility method.

In order to make two controlled vocabularies compatible, each compatibility method presented below adds a single layer to the two (or more) participant controlled vocabularies. And this single layer limits the semantic flexibility required to facilitate subject access interoperability. It is proposed then, that the conceptual framework of a subject access interoperability mechanism consist of at least three additional semantic layers to foster the necessary semantic flexibility, not present with one layer.

To come to this conclusion the paper 1) present the taxonomy of compatibility methods 2) outline how, with the addition of two extra layers, a conceptual framework of a subject access interoperability mechanism could be built.

4. WORK RELEVANT TO SUBJECT ACCESS INTEROPERABILITY: METHODS OF CONTROLLED VOCABULARY COMPATIBILITY AND CONVERSION

A short but relevant list of work related to controlled vocabulary compatibility and conversion follows. This is not an exhaustive list. A thorough bibliography of compatibility methods from 1960- 1995 is presented in Dahlberg (1996c).

4.1 Mapping

Mapping between two classification schemes is a matter of geometry. If the "Classification of Subjects...amounts to transforming the system of points marked out in a multi-dimensional space into a system of points along a line," (Ranganathan, 1967), then mapping one classification scheme to another is simply a matter of intersecting the lines. However that only matches classes to classes. Contexts in which the classes exist within their individual vocabularies, such as hierarchical inheritance or related information are not reflected in mapping. There is potential for information loss when one class of greater extension is mapped to a class of lesser extension; where extension is roughly defined as the number of entities (or range of entities) of the class, whereas the intension has for its measure the number of characteristics used in deriving it from the universe of subjects (Ranganathan, 1967). Philosophy has greater extension than Ethics. Ethics has a greater intension than Philosophy. Similarly, a class of greater intension mapped

onto a class of lesser intension loses information as well. The same is true if we think of an index as a line, except mapping in this case is matching a term with a term. Thus mapping to achieve subject access interoperability by intersecting two classes from two schemes does not work. An example of mapping is found on the Cataloging-in-Publication Data put out by the Library of Congress. There, the Dewey Decimal Classification and the Library of Congress Classification intersect. Mapping is limited in its semantic power. It has no intervening layer to control meaning, relationships, etc. It is clear that at least one layer behind these two classification schemes could facilitate the retention of information, and build its semantic power. By adding this layer mapping becomes switching.

4.2 Switching

Switching is mapping via a third component – a switching language. Lancaster (1986) says a switching language "can be used to convert from any one vocabulary to another... Here X represents the switching language," [in Figure 1].

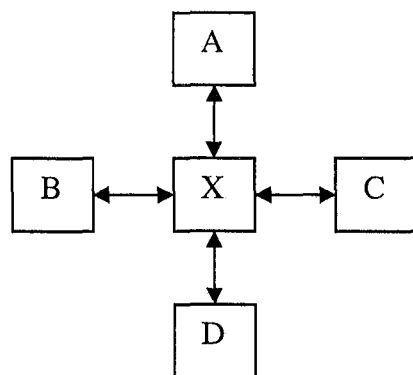


Figure 1. A schematic of a switching language

"Within an on-line network, this would allow a user of B to interrogate the data bases of A, C, and D, as well as B, using only B's vocabulary. Through the switching language, this [B's vocabulary] will be converted to any one of the other vocabularies," (Lancaster, 1986).

Lancaster lists ways in which the structure of subject access systems would confound any seamless switching. The problems inherent in switching between two vocabularies are 1) overlap of subject matter, 2) specificity, 3) degree of pre-coordination, and finally 4) hierarchical, synonymous, and other relationship structure (Lancaster, 1986). It is X, a single added layer (the switching language), that must control the different variables of meaning, relationships, extension, and intension. The problems highlighted by Lancaster can be effectively addressed via a number of layers, each with its own method of controlling one or two aspects of overlap of subject matter, specificity, degree of pre-coordination, and relationship structure. Implementing these layers, each working in coordination with the other, transforms the state of controlled vocabulary compatibility, a method that uses a single additional layer, into subject access interoperability, a method of multilayered flexibility. Lancaster's problems with switching show the first instance of where a multilayered conceptual framework would prove beneficial to facilitating subject access interoperability.

4.2.1 Information Coding Classification (ICC)

Ingetraut Dahlberg (1996a) proposed the Information Coding Classification (ICC), a universal classification scheme, as a switching language. The process of using ICC as a switching language among other universal schemes (UDC, DDC, etc.) is a three step process. "The *first step* is to correlate the classes of one classification system after the other with the subject groups of ICC. ... Doing this, it becomes obvious at which positions in the correlated systems there are gaps or only partial equivalences corresponding to the concept in question. For problems of this sort, a series of symbols taken from the mathematical symbolization of languages (such as <, >) were introduced," (Dahlberg, 1996a). The ICC would be X in Figure 1. above, and the universal schemes would be A, B, C, D.

Dahlberg solves the issues related to overlap of subject matter, specificity, degree of pre-coordination, and finally hierarchical, synonymous, and other relationship structure (Lancaster, 1986) with a.) the use of mathematical symbolization, and b.) in filling in the gaps, as she says in her third step. "The *third step* will consist in ironing out inconsistencies in the systems under comparison: filling in the gaps and seeing to it, that they receive correct symbolization," (Dahlberg, 1996a). The second step is a construction of lists of correspondences.

The ICC is a universal classification scheme. It has its own vocabulary and its own structure. And though it is not based on academic disciplines, but rather on the theory of integrated levels (Dahlberg, 1996a), it is a single language, and must reconcile conceptual discrepancy with mathematical symbolization. This implies, that the ICC does not record the most precise or more compound concept, but rather drops the user into a near neighbor.

In general, the ICC, in its design, offers solutions to the problems of disciplinarity and subject classification. These solutions are outlined in Dahlberg (1996). However, the ICC does not express all the levels of intension and extension required for an interoperable system, as evidenced by the use of mathematical symbolization. Yet it seems that the scheme of the ICC could serve as a component of a subject access interoperability mechanism. It could perhaps provide a relationship structure necessary in recognizing different semantic layers. But because it is a prescriptive classification, universal in scope, and operates on a single layer, it is doubtful that it, by itself, can support the multilayered conceptual framework of a subject access interoperability mechanism.

4.2.2 Broad System of Ordering (BSO)

The Broad System of Ordering is "a coding and ordering system for subject indication," (Coates et al., 1978). It was constructed "for the purpose of interconnection of information systems in the framework of the UNISIST programme, [to] design and develop a broad subject-oriented scheme, which will serve as a switching mechanism between information systems and services using diverse indexing/retrieval languages," (Coates et al., 1978). The BSO is very broad. The authors relate in their introduction that the subjects contained in the BSO could be more finely discriminated. They define the warrant for the BSO by saying, "if an independent[ly] organized information source devoted exclusively to a given subject is identified, then that subject should have a specific BSO code. If the notion of 'organised information source' is confined to

secondary sources such as abstracts, reviews, or indexes, BSO most certainly meets this criterion at the present time," (Coates et al., 1978).

By its scope and warrant the BSO does not risk losing information or many relationships when employed as a switching language. However, there is not much gained with its use. The control over constituent vocabularies cannot be extended to a full degree of specificity. What it gains in information retention, it loses in specificity. Were it multilayered, then perhaps it could gain control without losing information.

The BSO, like the ICC, offers insights into problems and solutions of controlling the compatibility between two or more controlled vocabularies. However, the BSO, also like the ICC, is a prescriptive classification, universal in scope that operates with a single layer of control. Alone, it cannot offer a multilayered flexible framework for a subject access interoperability mechanism.

4.2.3 Switching and inherent classification

As illustrated from the above examples, switching via a universal switching language, presumes an *inherent* classification of concepts. Inherent classification, like mapping, conflates the layers of meaning, distinguished by the relationships, extension and intension of controlled vocabularies. Concepts are separate units distinct from their terms or classes associated with them. Separating and recording layers is important for context. History, as a concept, can be classed anywhere – in the social sciences or humanities. History as a concept can be related to any subject, and any other class. Decisions regarding the placement of History within a controlled vocabulary are based on context and are an important element in information retrieval for each user group. In order to preserve meaning and facilitate coextensive subject access interoperability, concepts must not suffer from *inherent classification*. Thus, they must be considered in and of themselves, emancipated from the inherent meaning thrust on them by a single switching language. Then they can be considered in the context of many layers, facilitating a flexible, yet meaningful, subject access interoperability model.

4.3 Supra-thesaurus

H. H. Neville (1970) constructs a process for reconciling vocabulary differences between three thesauri in a common subject area. He identified the discrepancy between terms and concepts when approaching controlled vocabulary compatibility. He says, "[t]he concepts themselves will often not correspond as between one thesaurus and another: a specific concept in one thesaurus may be covered in another thesaurus by a broader concept; some concepts in any thesaurus may not be provided for at all in another," (Neville, 1970). He accounts for this discrepancy between multiple thesauri by coding, with numbers, concepts derived from terms in different thesauri. The process of analyzing and coding these concepts is called "reconciliation", (Neville, 1970). Each code represents a concept that is coextensive with a term, a combination of terms, or a missing term in participating thesauri. The process of building this list of codes, called the supra-thesaurus, is an iterative concept analysis of participating thesauri. Neville (1970) outlines the process as such. "It is supposed that there is a group of thesauri A, B, C, D to be reconciled. One thesaurus, say A, is taken as the 'source thesaurus' and all its keywords are taken into the

joint system by being given code numbers, and reconciled with all the other thesauri. This involves considering the type of reconciliation method required for each keyword, inserting certain additions to the source thesaurus and to the other thesauri, and compiling for each thesaurus a key to code numbers. This may be referred to as 'Round 1'. This is followed by Round 2, in which thesaurus B is taken as the source thesaurus, and all those keywords in it which were not dealt with in Round 1 are now reconciled in the same way with all the other thesauri, including A. Further rounds deal with progressively fewer residual keywords in the remaining thesauri," (Neville, 1970).

A supra-thesaurus is a switching language in its functionality. However, it is not like the ICC or the BSO. The latter are designed to switch any controlled vocabulary with any other controlled vocabulary. They are universal in scope and design (the BSO is decidedly general and states this). Each of these languages is a system that exists before the concepts that they are to switch between exist. The supra-thesaurus does not exist before the constituent thesauri are reviewed. It is generated as needed. Whereas the classes of the BSO and ICC are prescriptive (they prescribe – or are written before – the classes that can be used in switching), the supra-thesaurus is descriptive. It describes the concepts as they appear in constituent controlled vocabularies.

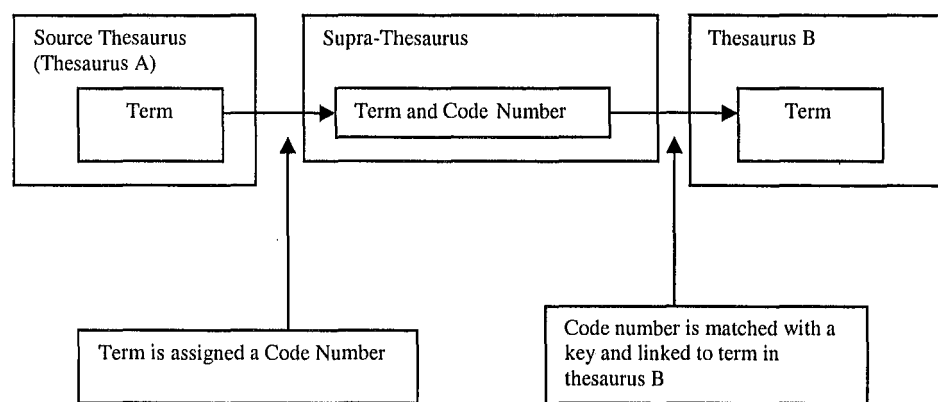


Figure 2. A schematic of a supra-thesaurus

4.3.1 *Supra-thesaurus and conceptual warrant*

The supra-thesaurus, constructed out of thesaurus reconciliation, is a collection of concepts that exist between the constituent thesauri. In effect, the supra-thesaurus is a record of *conceptual warrant* derived from these controlled vocabularies. In contrast to literary warrant, conceptual warrant is not based on the body of literature in a collection, but rather the collection of concepts from a controlled vocabulary. Both literary and conceptual warrant play a role in subject access interoperability. Each must be accounted for in an individual semantic layer in the conceptual framework of a subject access interoperability mechanism. Conceptual warrant also addresses some of the issues outlined by Lancaster above. The degree to which different controlled vocabularies overlap in subject matter, or differentiate in their level of specificity is determined by conceptual warrant. Thus it seems necessary to acknowledge conceptual warrant in order to address Lancaster's problems with switching languages.

4.4 Universal Source Thesaurus

Dagobert Soergel (1974) envisions a Universal Source Thesaurus (UST) to be a "cumulative thesaurus, using a great number of existing indexing languages and thesauri as input and precisely storing each bit of information contained in them. Therefore, a UST could be used as a data base of terms," (Soergel, 1974). This UST would be very detailed and complete. Because of its detailed and cumulative nature, the UST reflects the idea of conceptual warrant. However, Broader Term (BT), Narrower Term (NT), and Related Term (RT) relationships would be distinguished in the UST. It would seem to be counterintuitive to infuse a compatibility tool with a prescribed syndetic structure, especially when structure is considered a boundary to compatibility (Lancaster, 1986). Further, the user community of each controlled vocabulary would have their own interpretation of relationship structure. Each community could ask if the wolves belong in the same array as dogs. In answer to the need for more than one hierarchical structure Soergel states, "[a]ny relationship contained in any of the sources [contributing to a UST] (or suggested by any serious user) would be included, (Soergel, 1974). This is a necessary addition to a switching language according to the fourth problem outlined by Lancaster above (1986).

In order to make indexing languages A and B work together in the UST, two conversion tables must be constructed. These conversion tables lay out terms found in each indexing language, and define their relationships. The UST with its conversion tables, taken as a whole, is a switching language. And the UST with its conversion tables is designed by Soergel to be used as one. As a result, the UST suffers from the same problems outlined above. A single layer is used to control multiple problems.

However the overall structure and design of a UST, including Soergel's proposed management plan for it, are very helpful in planning the design and component structure of subject access interoperability. The work Soergel has done on the UST and compatibility will be invaluable in shaping components of a subject access interoperability mechanism. For example, tags to describe collection sizes (Soergel, 1974) could be used in a coordinated multilayered subject access interoperability mechanism.

4.5 Problems with controlled vocabulary compatibility methods

Not including Mapping, two main problems with the aforementioned controlled vocabulary compatibility methods (the switching languages) exist: 1) the type of control used to facilitate subject access and 2) the limited number of layers expressed in the switching languages.

The first problem is straightforward. Each of these compatibility methods uses a one dimensional syndetic structure in its switching language. The relationship structure is inherent in the switching language which inhibits the resolution of any of Lancaster's problems. If the switching language used to facilitate compatibility between two controlled vocabularies is rigidly structured, then the amount of information loss is great, if for the sole reason that Broader Term/Narrower Term relationships skew the ontology of a term (its overlap of subject matter, specificity, if not degree of pre-coordination) not just its relationship structure. In order to construct an interoperable environment for subject access, meaning and structure must be

accommodated more effectively. In the conceptual framework of a subject access interoperability mechanism, control is exercised on different layers, which facilitates control and flexibility.

Possessing a single layer is the second problem inherent in the compatibility methods listed above. Each of them, Mapping included, exists in too flat of a structure to express different dimensions of control. In order to solve Lancaster's problems with switching languages, each problem must be separated from the others and solved by mechanisms and methods unique to its dimension. Having a single layer or switching language is not enough.

5. LAYERS IN THE CONCEPTUAL FRAMEWORK FOR SUBJECT ACCESS INTEROPERABILITY

Each of these problems Lancaster (1986) identifies in switching languages: 1) overlap of subject matter, 2) specificity, 3) degree of pre-coordination, and 4) hierarchical, synonymous, and other relationship structure, stems from trying to use a single layer to facilitate compatibility between two controlled vocabularies. Each problem, including problems beyond those outlined above, must be dealt with individually. The conceptual framework for subject access interoperability presented below is multilayered, and as such addresses all of the problems inherent in providing full subject access across controlled vocabularies.

5.1 Concepts

In the above compatibility methods, each of the switching languages identified the concept behind the terms of the constituent controlled vocabularies. In order to preserve meaning during switching, each switching language used must be at least as precise as the constituent controlled vocabularies. Precision in this case refers to any level of granularity defined by the terms in the constituent controlled vocabularies. Thus a compound concept (formed of perhaps many different concepts) must be recorded in the switching language if it appears as a term in a constituent controlled vocabulary. The idea of a concept is, at any level of granularity, the desideratum of information retrieval (Soergel, 1974). The concept, because of its primacy in the function of information retrieval, is the focus of at least one layer in a subject access interoperability mechanism. Further, as mentioned in 5.4, the concept layer will help address Lancaster's problems with switching languages. Thus the *concept* is defined in subject access interoperability as: 1) the desideratum of information retrieval, 2) an individual unit of knowledge in a controlled vocabulary, 3) the potential mechanism for precision in information retrieval, and 4) a constituent of a subject.

5.2 Subjects

A *subject* is another layer in a mechanism for subject access interoperability. Lancaster's problem with subject matter overlap and specificity are concerns for the subject layer. A subject is 1) the desideratum of a literature review 2a) "[a]n organized or systematized body of ideas, whose extension and intension are likely to fall coherently within the field of interest and comfortably within the intellectual competence and the field of inevitable specialization of a normal person," (Ranganathan, 1967), 2b) "a formal system of teaching and research, societies at the international level devoted to the subject and practice of the art, learned and popular journals

publishing research... ", (Foskett, 1991), 3) the potential mechanism for recall in information retrieval, 4) made up of concepts, either singly or in combination. The subject layer does not enforce a hierarchical structure. It tracks terms for subjects and identifies if one subject is related to another. Other layers of the subject access interoperability mechanism describe the interrelationships between subjects, these include classes and participant schemes provided by participant classificationists.

The two layers of *concepts* and *subjects* allow the constituent controlled vocabularies to express the overlap of subject matter and level of specificity. The level of control exercised over these two layers (subjects and concepts) is a matter of policy, interpretation of conceptual warrant and literary warrant, and the nature of the user groups utilizing subject access interoperability. And each layer will require a different level of control. The mechanism for subject access interoperability, as it is envisioned here, will be guided by a distributed network of classificationists, contributing their knowledge of their user groups, their discourse community, and their controlled vocabularies to the subject access interoperability mechanism. This grants control of policy, conceptual warrant, literary warrant, and relevant retrieval into the hands of information professionals who work with these users. These classificationists are called *participant classificationists*.

5.3 Classes

The third semantic layer in a subject access interoperability mechanism is the *class* layer. This layer describes the hierarchical relationships of concepts and subjects. The class layer is a subject classification scheme that is fully faceted and employs a postulate based citation order. This citation order can be constructed in a dynamic way in a networked environment. Control over the extension and intension of facets is recorded in the whole subject access interoperability mechanism, but the display of this information is controlled by participant classificationists.

5.4 Purpose of concepts, subjects, and classes

The purpose of these three semantic layers is to disentangle the layers of problems inherent in switching languages (Lancaster, 1986). If each of the problems outlined by Lancaster (1986) can be isolated and a framework for each problem can be built, then subject access interoperability is a state within our grasp. If through the concept layer, a participant classificationist, can isolate the object of a user's query, and if by subjects, the participant classificationist can place that query in a body of literature (warrant), and if through the class layer, the participant classificationist can place the retrieved set into potentially useful relationships, then the problems of subject overlap, specificity, degree of pre-coordination, and relationship structure (Lancaster, 1986) are resolved.

6. PART OF A WHOLE

Concepts, subjects, and classes form only one part, the semantic layer, of the multilayered conceptual framework for subject access interoperability. It is envisioned that there will need to be at least three more types of layers, each with distinct component parts, in order to flexibly control subject access between different controlled vocabularies in the networked environment. The issues outlined above were a direct answer to Lancaster's problems with the semantics of

switching languages. How do information professionals and researchers control overlap of subject matter, specificity, degree of pre-coordination, and relationship structures in switching languages? The answer proposed here is to stratify those semantic issues across at least three layers (concepts, subject and classes) on the outset, to allow a flexible control of each of the dimensions represented by those layers.

7. SUMMARY

Creating a state of subject access interoperability requires a negotiation between control and flexibility. The ideal state of subject access interoperability retains meaning and structure that can be interpreted by users in the networked environment, and this is done by disentangling the semantic layers involved in controlled vocabularies. This paper proposes that a single layer, found in many switching languages, is not sufficient to reconcile problems with 1) overlap of subject matter, 2) specificity, 3) degree of pre-coordination, or 4) hierarchical, synonymous, and other relationship structure. In order to address these problems a multilayered semantic layer consisting of at least a concept layer, a subject layer, and a class layer is proposed. This act of semantic disentanglement establishes a flexible control of each layer, rather than a more rigid scope of control provided by the single layer of the switching languages mentioned above.

WORKS CITED

- Coates, E., Lloyd, G., and Simandl, D. (1978). BSO: Broad system of ordering, schedule and index. (3rd rev.) The Hague, Paris: Fédération Internationale de Documentation; UNESCO.
- Dahlberg, I. (1996a). Library catalogs in the internet: switching for future subject access. *Advances in knowledge organization*, 5, 155-164.
- Dahlberg, I. (1996b). The compatibility guidelines – a re-evaluation. In *Compatibility and Integration of Order Systems (Research Seminary Proceedings of the TIP/ISKO Meeting, Warsaw, 13-15 September, 1995)*. Warsaw Wydawnictwo SBP.
- Dahlberg, I. (1996c). *Compatibility and Integration of Order Systems 1960-1995: an annotated bibliography*. *Compatibility and Integration of Order Systems (Research Seminar Proceedings of the TIP/ISKO Meeting, Warsaw, 13-15 September, 1995)*. Warsaw Wydawnictwo SBP.
- Doerr, M. (2001). Semantic Problems of Thesaurus Mapping. *Journal of Digital Information*. 1(8). [On-line] <http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr>
- Foskett, D. (1991). Concerning general and special classifications. *International Classification*. 18(2) 87-91
- Koch, T. and Day, M., et al. (1997). Role of classification schemes in internet resource description and discovery. [On-line] <http://www.ukoln.ac.uk/metadata/desire/classification/>
- Lancaster, F. W. (1986) *Vocabulary control for information retrieval*. (2nd ed.) Arlington, VA: Information Resources Press.

Proceedings of the 12th ASIS&T SIG/CR Classification Research Workshop

Neville, H. H. (1970). Feasibility study of a scheme for reconciling thesauri covering a common subject. *Journal of Documentation*, 26(4) 313-336.

Ranganathan, S. R., & Gopinath, M. A. (1967). *Prolegomena to library classification* (3rd , ed.). London: Asia Publishing House.

Soergel, D. (1974). *Indexing languages and thesauri: construction and maintenance*. Los Angeles: Melville Publishing Co.

