Nancy J. Williamson  
Faculty of Information Studies, University of Toronto  

Deriving a Thesaurus from a Restructured UDC  

Abstract: The derivation of a thesaurus from a new schedule for UDC Class 61 Medical Sciences which has been restructured into a faceted classification system using the framework provided by the Bliss Bibliographic Classification. The resulting thesaurus is intended to serve as a tool for indexing and searching but will also be the index to the 61 class itself. The background for the research is briefly described. The sources and methods used to select the descriptors and define their relationships are discussed. Problems are identified and some solutions proposed.

1. Introduction  
This research is the third stage of "A Feasibility Study on the Restructuring of the Universal Decimal Classification into a Fully-Faceted Classification System" (McIlwaine and Williamson, 1994), described at the Third International ISKO Conference in 1994. The study is a pilot project in which the UDC Class 61 Medical Sciences is being reorganized and restructured using the facet framework of the Bliss Bibliographic Classification, second edition (BC2) as the basis for overall organization and content. Reports on progress are described in several articles (McIlwaine and Williamson, 1993; Williamson, 1994). The first stage, the reorganization of the topics, is nearing completion and portions of the restructured schedules were published in Extensions and Corrections to the UDC, 1995 (McIlwaine and Williamson, 1995). The second stage, the assignment of the notation, is in progress (Williamson, 1995) and the third stage, the derivation of a thesaurus from the restructured 61 schedules, is in its initial stages. The fourth and final stage will be the review and testing of the results.

2. Background and Previous Research  

Normally, information retrieval systems should provide for two fundamental approaches to searching—direct access through the use of verbal descriptors and browsing capability through the use of systematic knowledge structures. Together they supplement and complement each other. Classification systems and their indexes are tools which combine these two approaches, offering the potential for optimal aid in searching information systems. Moreover, the development of a systematic knowledge structure as the first stage in the preparation of a thesaurus is a sound approach to the construction process. It helps to clarify the parameters on the subject domain and acts as a mapping device for organizing the contents of the system. It is the derivation of a thesaurus by this process, that is the subject of this paper. In this context, it is assumed that the thesaurus will be used not only as a tool for indexing and searching in systems using the new UDC Medical Sciences class, but that it will also serve as an index to the 61 class itself.

While all large classification systems require indexes for effective and efficient use, the type of index varies with the system. For example, the Dewey Decimal Classification is well known for its "relative index" and Ranganathan developed "chain indexing" to support the Colon Classification. Nor is the use of a thesaurus as an index to a classification scheme a new idea. Over the past 25 years a number of thesauri have been compiled by deriving terms from classification schemes. In some cases a new classification has been created and the thesaurus
derived from it, while in others existing schemes have been used to derive the thesaurus. Typical examples are Thesaurofacet (Aitchison, 1967), BSI ROOT Thesaurus, and the International Thesaurus of Refugee Terminology (Aitchison, 1989). In some cases the resulting thesaurus has been an alphabetical thesaurus alone, while in others the classification used to derive the thesaurus has been incorporated into the system, becoming an index to the classified structure. This kind of a thesaurus is best generated using a faceted (as opposed to an enumerative) classification. Indeed the title Thesaurofacet itself aptly describes the format of such a system. A variation on this type of system has been referred to as a "Classaurus".

Many of the systems, developed in this way are a result of research carried out by the British Classification Research Group and others. Of particular interest in this study is the research carried out by Jean Aitchison (1986), and Gerhard Riesthuis and Steffi Bleidung (1990, 1991). They represent two different approaches to a similar problem. Riesthuis and Bleidung experimented with Class 314 Demography in the existing semi-enumerative UDC. Building on the fundamental premise that alphabetical and systematic systems are complementary to each other, the authors created two thesaurus displays. Both contained preferred terms displayed with their scope notes, Broader Terms, Narrower Terms and Related Terms and equivalence relationships. An additional category designated as "Use for Combinations" (UFC) was included to take care of terms which are precoordinated terms in UDC but needed to be factored for the thesaurus because the component terms have their origins in different hierarchies. This took care of the use of colon combinations which are an important feature of the UDC system. Containing the same data, the two thesauri were arranged differently at the primary level. One was organized alphabetically and the other was a systematic thesaurus (as opposed to a classification scheme) arranged by UDC class number. Findings revealed the weaknesses of the UDC system, among which were out-of-date and missing terminology, and the complexity of the enumerative system which hinders the efforts to convert into a concept based system such as a thesaurus. Nevertheless the authors concluded that, given the time and effort to revise the schedules, UDC might be converted to a thesaurus with two complementary parts—alphabetical and systematic.

Aitchison, a member of the CRG and an expert in thesaurus construction, has developed a number of thesauri of this type, some by first constructing the classification and then deriving the thesaurus; other by drawing on an existing classification systems. Out of her experience comes her experimentation with and use of the Bliss Bibliographic Classification second edition as the basis. The DS/HSS-Data Thesaurus (Aitchison, 1993) provides controlled terms and Bliss class numbers for indexing and searching and the ECOT Thesauri (Aitchison, 1984) uses software designed for the British Standards Institution ROOT Thesaurus to generate the alphabetical display from the systematic display derived from the Bliss schedules (Aitchison, 1986, 160). Normally, the pattern here has been to include both the classification scheme and the thesaurus. Not unexpectedly, Aitchison demonstrates that the best basis for such work is a well structured faceted system and opposed to a enumerative scheme. Also from her research Aitchison has identified problems and developed some rules and policies for conversion. Together with the ISO 2788 Guidelines for the Establishment and Development of Monolingual Thesauri (ISO, 1986) and Thesaurus Construction: A Practical Manual (Aitchison and Gilchrist, 1987), the work of these researchers forms the basis for the current project. The two pieces of research are particularly apropos to this project, since together they deal with the two classification systems being used.

3. The Restructured Class 61

In terms of its structure the "new" UDC Class 61 follows the BC2 Class H schedule for "Health Sciences. Medical Sciences". The result is a complete reorganization of UDC, as the BC2
structure emphasizes collocation by systems and organs of the body (e.g., Cardiovascular System, Digestive System, etc.) as opposed to collocation by characteristics such as Anatomy, Physiology, Clinical Medicine, etc. BC2 facet indicators are part of the schedules and the captions are based on simple concepts. Of the two systems, BC2 is a much more in-depth classification and much richer and more up-to-date in terms of vocabulary than the existing UDC. As a result the present UDC 61 is subsumed under the BC2 umbrella with some characteristics of UDC being retained. For the time being, this restructured class must be used in conjunction with other schedules in the existing UDC system. Hence the same kind of notation is being used and the new 61 schedule will be "accommodated" to the existing system where necessary, to permit cross referencing and the use of general auxiliary subdivisions from the main UDC schedules. All relevant vocabulary from both systems is being retained. However, since in general, the goal for UDC is a Standard Edition approximately equal in size to the UDC International Medium Edition, English Text, some very precise topics will be grouped together as "including notes" and subsumed under the appropriate caption as exemplified in Figure 1.

Immunological deficiency syndromes, immunodeficiency

Including Antibody deficiency syndrome.
Aldrich syndrome. Wiskott-Aldrich syndrome.
Thymic alymphoplasia. DiGeorge syndrome

Fig. 1: Including Note

Because the new 61 class is a faceted system, insofar as possible topics are given a primary location and the use of the UDC "colon combination" feature will be retained as a tool for number building. Using this device, concepts will be imported from their primary location to other parts of the system as needed. For example under "Nutritional disorders, dietary disorders, malnutrition" at "Food poisoning, (Higher plants)" the colon combination will be used to import the number for plants from class 58 to be added to the basic notation.

"Alternative" locations (with preferred location indicated) for numerous topics are a significant feature of BC2. Where appropriate some of these will be retained. Where this happens the thesaurus will need to accommodate both locations. The many auxiliary tables in UDC 61 are to be amalgamated, insofar as possible. Since most of the 61 auxiliary subdivision topics are located in the main part of the schedule in BC2, the number and size of the auxiliary tables in 61 can be much reduced.

4. Source of Terminology for the Thesaurus

Terminology for the thesaurus will be gathered primarily from the classification topics. Two problems must be addressed—the choice of preferred terms as descriptors and the determination of the form of those descriptors when removal from their context in the classification renders them meaningless or ambiguous. Frequently, multiple terms represent the same topic. Which term becomes the preferred term? BC2 tends to string terms together the captions separated by commas. A particular caption may contain multiple scientific terms or a mix of scientific and "lay" terms. Frequently, but not always, in BC2 the scientific term precedes the non-scientific term. While the choice preferred term should be based on a policy which will best suit users, for the sake of expediency, the first term in each string will become the preferred term in this thesaurus and "use" references will be created to accommodate the other terms. Figure 2 illustrates two instances of choice of preferred terms and the resulting references.
In general, synonymous terms will not be too difficult to deal with although they it must be certain that the terms are indeed synonymous or quasi-synonymous in nature. Much more difficult and time consuming is the task of providing context for terms which require qualification in order to be correctly understood. These terms are of two types. As illustrated in Figure 3, some terms are "incomplete" as they appear in the classification schedules and require the context of superordinate terms if they are be unambiguous when displayed the thesaurus. To solve this problem some terms need to be "precoordinated" for inclusion in the thesaurus.

As can be seen from Figure 3, the resulting descriptors may, in some cases, become cumbersome. Moreover, some of them will violate the ISO Guidelines for compound terms and factoring. Nevertheless this kind of precoordination will be necessary if the descriptors are to be understandable out of context of the systematic display. Such problems occur more often at the upper levels of a particular topic than they do under more precise topics located at lower levels in the hierarchies.

A second terminological problem is one in which a general term appears in a location in the schedules which is not its primary location. For example, terms which describe the properties of a subject may come from completely different domains. This is well illustrated at the most general level in "anatomy and physiology" where in the facet (Properties, etc.) such terms as "ratio,"
"periodicity," "rhythm," "dimensions," and "shape," can be found. In the appropriate situation all of these terms have relevance in this domain and locations must be preserved for them. Nevertheless when removed from context these terms take on a completely different meaning. If the thesaurus is to be an index to the classification, these terms must be accommodated in the thesaurus. While at the time of writing it is not clear how these will be handled, two approaches seem possible. One solution would be to retain only the facet indicators in the schedule and depend on special auxiliary tables to provide for the topics themselves. An alternative to this is to "qualify" the terms in the thesaurus. For example: "Periodicity (anatomy and physiology)," "Shape (anatomy and physiology)." This would be compatible with the standards and would preserve the classified display.

5. Source of Thesaural Relationships

The facet framework being used for the classification provides for a well organized schedule and the division of topics by facet and subfacet labelled with facet indicators clarifies the characteristics of division and makes it fairly easy to recognize relationships between terms and to extract them for use in the thesaurus. The reconstructed 61 schedule will permit the extraction of the traditional thesaural relationships—equivalence (Use/UF), hierarchical (BT/NT) and associative (RT/RT). In addition, scope notes and definitions from the classification will be used in the thesaurus. As discussed above, the equivalence relationships will be easily identified from the captions in the classification schedules. Figure 4 illustrates this.

Classification: Visual aphasia, alexia, word blindness

Thesaurus: Alexia Use Visual aphasia

Visual aphasia
UF Alexia
Word blindness
Word blindness Use Visual aphasia

Fig. 4: Equivalence Relationship

Hierarchical relationships are of two types, generic relationships and whole-part hierarchical relationship. The facet indicators are of immense help in identifying this particular relationship as there are number of "(By type)" and similar facet indicators. An example appears in Figure 5.

The hierarchical whole-part relationship covers a limited number of situations where the name of the part implies the name of its possessing whole is any context. Medicine is an excellent example of this kind of relationship, because parts, systems and organs of the body qualify for this type of relationship. For example "Cardiovascular System (Parts special to the heart)" includes the pericardium, ventricles, septum valves, blood vessels, etc. These would all be related to the whole in a BT/NT relationship. Polyhierarchical relationships are seldom made clearly explicit in Class H in BC2, primarily because of the faceted nature of the system. There are a few cases where compound heading appear but these are at the very
Classification: Parasites, parasitism, medical parasitology (general)

(Types of parasites)
Ectoparasites
Endoparasites

Thesaurus: Ectoparasites
BT Parasites
Endoparasites
BT Parasites
Medical parasitology (general)
Use Parasites
Parasites
UF Parasitism
Medical parasitology (general)
NT Ectoparasites
Endoparasites
Parasitism
Use Parasites

Fig. 5: Generic Relationships

top of the class for example the multi concept "Anatomy and Physiology" is used but only to provide for work which may deal with the two topics together. The system leans toward one topic or the other.

Classification

Cells
(Physiology) Cytophysiology
(Biochemistry) Cell chemistry
(Special physiological processes)
...
Motility
Aggregation
(Development and growth)
Change
Regeneration

Thesaurus

Cells
RT Cytophysiology
Cytophysiology
NT Cell aggregation
Cell change
Cell chemistry
Cell regeneration
RT Cells

Fig. 6: Associative and Hierarchical Relationships

Associative relationships are numerous in the new Class 61. Properties, processes, actions and their agents, actions and the products of the actions, among others are prevalent in the Medical Sciences. Typical examples are Respiration RT Absorption; Kidneys RT Transplantation; and Heart RT Open heart surgery; Determining the associative relationships is greatly facilitated by the presence of facet indicators such as (By action); (Surgery); (By chemical process): (Disorders by cause) etc.

The various thesaural relationships are present in the classification scheme and transferable to the thesaurus. However, the transfer cannot take place without careful thought at each step and in terms of each descriptor. This is particularly true of the BT/NT relationships. Some hierarchies are very long indeed and considerable patience is required to determine exactly which descriptors
are subordinate and superordinate to each other. This partly due to the physical nature of the schedules as well as to the length of the hierarchies.

6. Problems

Deriving a thesaurus from this structure is not without its problems. Some of these are a result of dealing with two quite different classification systems. Also, the fact that this project is focused on only one discipline which exists in the context of a larger system. As Aitchison (1986, 166) indicates, there are gaps in the coverage so that it is not always possible to make all of the necessary links with other descriptors. If this is true because all of the BC2 schedules have not yet been published, it is even more so with the new UDC 61. As a result of the great depth of detail incorporated into the new schedule through the use of the BC2 framework, there are many terms, particularly in other classes which may be present eventually in BC2 that are not present in other UDC classes. For example, there is detailed coverage of Biochemistry in Medical Sciences which draws heavily on Biochemistry in Science in BC2. However, this new schedule, while based on BC2 must be used in the context of the existing UDC. Repetition of concepts in many places may also add to the size and complexity of the thesaurus. However, alternate locations may not be a great problem because the enumeration of topics only takes place in the "recommended" location. The user of the classification scheme is left to develop the alternate. Another problem arises in dealing with the emphasis on parts, organs and systems of the body. Some topics are scattered in many places. A typical example is "Neoplasms" which go with the site of the disease. Thus the term neoplasms turns up in many places in the schedules all of which will need to be included in the thesaurus/index. In this context it is interesting to note that in the existing BC2 index there is only one entry for "neoplasms" which leads to the general section on Diseases. There a general directive can be found leading to "parts, organs and systems of the body" without more specific directions.

to the precise location in the various sites.

7. Conclusion

This is a fascinating undertaking which, with time and patience, may be accomplished. However, it is a compromise at best. The results should give a more logical system which will be better suited to use in online systems than its predecessor, but the ideal situation would be a completely new classification system.

References


Aslib.


