

The Universal Decimal Classification: Research to Determine the Feasibility of Restructuring UDC into a Fully-Faceted System

Nancy J. Williamson

Faculty of Information Studies

University of Toronto

140 St. George St.

Toronto, Ontario M5S 1A1

william@fis.utoronto.ca

Describes the nature and progress of experimental and exploratory research to determine the feasibility of restructuring the Universal Decimal Classification. The background leading to the research is described. The study uses the facet framework established in the Bliss Bibliographic Classification 2nd edition (BC2) Class H and the discipline being restructured is UDC Class 61 Medical Sciences. The background for the study is provided and the purpose and scope are stated. A methodology for carrying out a complete restructuring is being tested. Problems with the process are identified and solutions presented. The general framework proposed for the class is presented with examples of possible results from the restructuring process. There is a possibility that one of the by-products of the research may be a model for future new classification systems in special subject areas and that it may provide some insight into classification requirements for online systems of the future.

BACKGROUND

In spite of the introduction of computers and the advent of new methods of information retrieval, our general classification systems have continued to be designed, updated and used in much the same manner as they were 100 years ago. The Universal Decimal Classification, which will be 100 years old in 1995, is no exception to this rule. However, even though it has at least 100,000 users and is available in a number of languages, it is in need of drastic revision and updating. Changes in administrative and editorial status and a desire to determine ways and means of improving the utility of UDC in computerized databases and information systems have led to a project which, if successful, could result in a vastly different classification system from UDC as it presently exists. The desire for changes in UDC is not a new idea and this project is a further step on a continuum which began approximately ten years ago.

In recent years, UDC has been described as "a classification system in crisis". Crisis situations in varying degrees are constant to classification systems most of the time, but this is more true of UDC than some other systems. Cumbersome methods of revision and lack of financial support have plagued UDC for almost of all of its entire 100 years of history, yet it remains popular with its users because of its flexibility, the power of its synthesis and its availability in many languages and in varying degrees of depth of analysis — full, medium and abridged editions. For these reasons, during the past few years those responsible for its management and editing have been investigating an effective means of bringing UDC into the 21st century as an efficient tool for the online environment.

PROCEEDINGS OF THE 5th ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP

Over the past few years, a number of events have improved the administration, editing and production of the schedules and placed UDC on a better regulated financial footing. In the early 1980s the International Federation for Information and Documentation (FID) commissioned an external study of the management of UDC and its revision process. The study, undertaken by Alan Gilchrist, well known as a United Kingdom management consultant and expert in thesaurus construction, was submitted to FID in 1984. Following from this a limited-life UDC Management Board was established and this group produced a "Proposal for the Future Organization of UDC Management and Revision Structures". This Management Board has since been replaced by a UDC Consortium (Goedegebuure 1993) made up of representatives from the 5 leading publishers from Belgium, Japan, the Netherlands, Spain and the United Kingdom, who together with FID, jointly fund the operation of UDC and its Central Secretariat.

While the original Management Board was concerned with the overall management of UDC, in 1988 the Board established a limited-life Task Force on the UDC System Development. The Task Force's brief was to undertake a thorough examination of the UDC system itself and to make recommendations for its future. If the system appeared to have a viable future, the Task Force was instructed to suggest a development plan. An investigation of a sample of users from approximately 10 countries determined that the information community was strongly in favour of the scheme. As a result, proposals were made which were designed to improve UDC's currency and its suitability for machine retrieval. The Task Force submitted its report in 1990 and chief among the recommendations were the establishment of a machine-readable database of the UDC schedules, the complete revision of the system over a ten year period using a faceted structure, and the enhancement of UDC's usefulness in a computerized environment by adding an alphabetical thesaurus as the index. It was from these recommendations that this present project has evolved, approved and encouraged by the UDC Consortium.

PREVIOUS RESEARCH

Classification systems, particularly faceted systems, and thesauri have much in common in their systematic approach to the organization of knowledge. The research literature identifies at least three types of links between thesauri and classification systems. There are instances (Marosi 1969) in which independent classification systems and thesauri have been used to support each other in the retrieval process. Also, a considerable number of thesauri have been developed with fully integrated components of a classification and a thesaurus. *Thesaurofacet BSI ROOT Thesaurus* (1988) and any number of systems produced by the Classification Research Group fall into this category. Finally, several thesauri, including the *DHSS-DATA Thesaurus* and the *ECOT Thesaurus*, have been derived from classification systems, while the *BSI ROOT Thesaurus* itself has been used as a basis for other thesauri. Fundamental to these systems are the idea of the simple concept, the principles of facet analysis and equivalence, hierarchical and associative relationships among concepts. Moreover, basically enumerative systems, such as the Dewey Decimal Classification, are strongly influenced by facet principles. UDC as it currently exists might be described as 'semi-faceted' but perhaps more correctly as 'semi-enumerative'. Expressed in another way, its underlying principles make it an "aspect" classification and it is "hierarchical" and "synthetic" in nature. (McIlwaine 1993) permitting the pre-coordination of parts of a subject which are not enumerated in the schedules.

The close association between thesauri and classification was the basis for a suggestion that UDC should become the basis for a thesaurus (Reisthuis and Bliedung 1991). The authors identified problems related to structure and to gaps in the classification, but among their conclusions were recommendations that efforts be made to "try to make UDC more faceted than it is now" and that rules should be developed for deriving a thesaurus from UDC. An extension of this research was the idea (Williamson 1992) that UDC itself might be restructured and the restructured system be used to derive the thesaurus. Support for this idea was provided in research carried out by Jean Aitchison (1986) in which she used the *Bliss Bibliographic Classification*, 2nd ed. as a source of thesaurus terms and structure.

THE PROBLEM

There are at least three ways in which the UDC problem might be addressed. One approach might be to make minor adjustments to the present system, making a limited number of structural changes and reducing the multi-concept topics to simple concepts. A second Scenario might be to replace UDC with a new classification system constructed without regard for the existing system. A third possibility is to map UDC classes and subclasses onto a facet structure which already exists. This latter alternative appears to be the most viable because some of the ground work has already been accomplished, but it also assumes that a suitable classification system can be found which can be used in the mapping process. The *Bliss Bibliographic Classification* (BC2) second edition lends itself to this mapping role in several ways. It is a comprehensive fully-faceted system which is recognized as having a sound and logical structure. Its schedules are recent and relatively up to date. BC2 is well and clearly documented as to its nature and the rationale for its facet framework. It provides helpful definitions, a clear rationale for facet indicators and suggestions for alternatives in the organization of concepts. Also BC2 has already been tested as a mapping device, since it has been used as the basis for the derivation of several thesauri. However, its use in this role is not without its problems. With respect to this project, important considerations are the fact that its development and publication are still in progress and there are major differences in the general nature of the two systems — BC2 and UDC. Nevertheless if such an approach is to be tested, BC2 is the most viable system in existence for this purpose.

PURPOSE OF THE RESEARCH

This research has several purposes. The principal objective is to deal with the future of UDC — to determine the feasibility of a major restructuring of the whole UDC system. In this respect there are a number of questions to which answers will be sought. Is such a radical restructuring feasible? Useful? How much human effort is involved? Other questions are more technical in nature. What level of division would be required for a "standard" edition of UDC? What level of faceting is appropriate? And what are the notational requirements for the restructured system? A second purpose of the research is to determine a method by which the restructuring might be carried out efficiently. A model is needed which could be used for UDC and which might also be used in the future to construct new classification systems in special subject areas. In a broader sense, the principal investigators are hoping to provide some insight into what a classification system for the 21st century should look like? While the investigation is unlikely to provide a definitive answer to this latter question, it is one way of addressing the question.

PROCEEDINGS OF THE 5th ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP

SCOPE OF THE PROJECT

This project focuses on one discipline in the universe of knowledge. The facet framework in Class H in BC2 will be used to reorganize and restructure the Medical Sciences (Class 61) of UDC. Medicine has been chosen because it is one of the UDC classes which is most in need of revision. For purposes of the study the UDC Master Reference File (MRF), the machine-readable version of the classification, presently available from the UDC Consortium is being used. This version has been developed to approximately 70,000 records, the size of the International Medium Edition, English Text, 2nd edition published in 1994. A thesaurus will be derived from the restructured class and two bibliographic databases will be created for testing the system. One database will contain documents classified by the existing UDC and the second database will contain the same documents classified by the newly restructured system. Advice of UDC users and subject experts is being used in the course of the project.

The effectiveness of the retrieval and the document display using the reconstructed system will be evaluated. The final result will be a report with recommendations to the UDC Consortium on whether a complete restructuring of this nature should go forward and, if so, how any restructuring process should proceed.

METHODOLOGY

Since there is no precedent for this research, very careful planning was required and some preliminary work was needed to set up a methodology. As a result the research divides itself into 7 broad phases as follows:

1. Preliminary investigation
2. Establishment of a methodology and a trial run,
3. Complete restructuring of UDC Class 61
4. Development and application of a notation
5. Derivation of a thesaurus
6. Evaluation and review
7. Testing of the system and its application to documents

Recognizing the magnitude of the project and the vast differences between the two classification systems, preliminary investigation was necessary in order to identify problems which would be encountered and questions which need to be addressed from the beginning of the research. The two schemes differed greatly in structure. How should this be handled? They also differ in degree of division and depth of analysis? What level of detail should be the goal? Among other concerns were schedule order, citation order and notation.

To provide tentative answers to these and other questions, in the early part of 1993, a pilot study was carried out on sections of the UDC 61 and BC2 H classes (McIlwaine and Williamson 1993). As a first step the relevant parts of the two schemes were obtained in machine-readable form. The UDC MRF was already available in CDS/MICROISIS and BC2 Class H was scanned and both systems converted to WordPerfect 5.1 for each of searching. Then the overall structure of BC Class H was examined to identify a general framework for the whole restructured class and a detailed analysis was carried out on one specialty "Dentistry". Subject specialists, librarians and doctors were also consulted. Based on the findings, it was decided that *insofar as possible* the BC2

structure would be used, that all BC2 facets would be included and the BC2 schedule order would be incorporated into the restructured system.

From the preliminary investigation it was possible to establish tentative general framework for the restructured class and a step by step process for inserting UDC topics into the BC2 facets. The process is being applied in a trial run working topic by topic and moving from the top down. Specifically anatomy is being restructured first, followed by physiology, followed by the systems of the body. The steps of the process are as follows:

1. A rough sort of UDC topics into the BC2 framework;
2. Refinement of the rough sort and the creation of a new database;
3. Review, evaluation and readjustment of the process.

Insofar as possible, the rough sort is being carried out by executing the search command in WordPerfect and relocating each retrieved caption in the appropriate BC2 facet. Inevitably, there "orphan" terms from both schemes, facets which remain empty in BC2, and topics for which there is no obvious facet to which the UDC terms belong. Another problem is differences in terminology for the same concept between the two systems. After a second machine-search to pick up additional concepts not located in the first search, the "orphan" terms are set aside to be dealt with separately on an individual basis. These are considered manually and slotted into existing facets, put into a newly created facet, or set aside for later consideration and possible elimination. Empty facets are retained as a part of the overall framework for use in the future when new topics appear.

The second step in the restructuring is to examine and analyze the rough sort and to refine it to ensure that terms are placed in what appears to be the correct facet. At this stage the work will be supported by the use of tools from the discipline - medical dictionaries and structured tools such as the MeSH tree structures and the NLM Classification. Some terms will need to be repositioned and some will require breaking apart into simpler terms. At this point the goal is to make the structure as intellectually sound as possible. Among the questions which will require positive answers are:

Is this caption in an appropriate place in the scheme?

Should the principal location of this topic be elsewhere in the system and "imported" into this location to be synthesized with this topic as needed?

Should all or part of this term be part of the UDC auxiliaries rather than a main schedule topic?

Is this topic located in the correct facet?

Is this topic a multi-concept which needs to be broken down into its component parts?

At this stage also, MEDLINE will be used to determine, on the basis of literary warrant, the need for inclusion of more specialized terms in the scheme. When the "Trial Run" has been completed, the results will be reviewed and evaluated and experts and users consulted with a review to locating

PROCEEDINGS OF THE 5th ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP

problems which need to be addressed and adjusting the process before the full scale restructuring begins.

With, hopefully, what will be a tested, refined, and adjusted methodology the whole class will be restructured working from the top down. It is essential to work in this manner, as what happens in the higher levels of the class will affect the requirements for the later sections of the schedule to a considerable degree. To keep redundancy and repetition at a minimum, terms which have their primary position earlier in the schedules will not normally appear in secondary positions in later parts of the schedules but will be "imported" and linked by a colon (or other device) to a primary topic which appears elsewhere in the schedule. When the restructuring is completed, a full review and evaluation by experts and users will take place.

When the restructuring is complete, the notation will be added. As yet it is not clear what kind of notation this will be. However, it is clear that the notation presently in use in UDC will be decimated when the restructuring is complete. It has been suggested that the new notation be UDC-like. Among the essential requirements are that it must be able to handle the schedule order of the restructured system and it must be understandable across languages. The final step in the creation of the 61 schedules will be to derive a thesaurus from the classification scheme. This will be done in a manner similar to the work carried out by Aitchison (1986).

The final phase of the project will be the review and testing of the system from two points of view. First of all, the classification will be critiqued and evaluated for its vocabulary, structure and content and the thesaurus will be measured against the ISO 2788 guidelines for monolingual thesauri. Secondly, the classification system will be tested on a random sample of documents. The sample will be chosen from a collection in which they have previously been classified according to the present UDC system. These same documents will then be classified according to the restructured version and a comparison made. A precise methodology for the testing has yet to be worked out, but among the criteria for evaluation will be useful order, logical order, hospitality of the system, and ease of use.

Based on findings in the restructuring process, and final tests, a report will be prepared for the UDC Consortium with recommendations on the way in which future revisions of UDC should proceed.

FRAMEWORK OF THE PROPOSED UDC CLASS 61

Before the process of restructuring could begin, it was necessary to make some decisions about the general overall structure for the class. Based on an analysis and general comparison of the two schemes and discussion with subject specialists, librarians and doctors it was decided that the general framework outlined below should be used. However, this decision had to be made in the light of a number of problems. The general arrangement of the two schemes is, in principle, not greatly different. Both begin with Human Biology and follow this with the specialties of Medicine. Bliss includes Anthropology in Class H with a number of other topics such as Genetics, while UDC locates it in Biology at 57. Similarly, Embryology is placed in the early part of the class well ahead of Medicine per se, unlike UDC. However, the major difference between the two schemes is in the citation order. Bliss treats the parts of the body as the leading category and subsumes all other aspects to it, so that a part of the body is subdivided by its anatomy, physiology, genetics, etc.,

followed by its diseases, pathology and so on. UDC, on the other hand, treats anatomy; physiology and pathology as three distinct subclasses. Although notationally there is a link through the use of parallel subdivision, the whole approach to the discipline is reversed. Based on modern medical opinion it seemed best to group by systems of the body (the Bliss order) rather than by process. Also, as a result of the preliminary investigation and the consultation with subject specialists, it was decided that if the task is to be performed satisfactorily, it is essential to begin at the beginning of the BC Class. It is important to begin with the fundamental disciplines such as embryology and genetics before proceeding to the specialties of medical diagnosis and treatment. Based on preliminary findings the tentative general framework for the class will be as follows.

Proposed Framework for Medical Sciences

- a) Pre-clinical Medicine
 - Human Biology (General)
 - Anatomy
 - Physiology
 - Biochemistry
- b) Developmental and Cell Biology
 - Cell Development
 - Human Embryology
 - Genetics
 - Haematology
 - Immunology
- c) Parts, Organs, Systems of the Body
(Including Diseases, Surgery, etc.)
 - Locomotor System, Musculo-Skeletal System
 - Cardiovascular System
 - Nervous System
 - Glandular System
 - Respiratory System
 - Digestive System
 - Urogenital System
- d) Public Health and Health Care
(Including hospital care, nursing, etc.)
- e) Systems, Schools of Therapy

The impact of this proposed framework on the present organization of UDC can be seen in the following:

PROCEEDINGS OF THE 5th ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP

UDC Reorganized

1.	Pre-clinical Medicine	
	Human Biology	57
	Anatomy	611
	Physiology	612
	Biochemistry	551; 577
2.	Developmental and Cell Biology	
	Cell Development	612.041.3
	Human Embryology	611.013
	Genetics	575
	Haematology	612; 615; 616
	Immunology	577; 615
3.	Parts, Organs, Systems	
	Each combines:	
	Anatomy	611
	Physiology	612
	Toxicology, therapeutics	615
	Pathology, Clinical Medicine	616
	Surgery, Orthopaedics, etc..	617
4.	Public Health and Health Care	
	Hygiene. Personal Health	613
	Public Health and Hygiene	614
	Preventive Medicine	614
	Nursing Care	616.083
	etc.	
5.	Systems, Schools of Therapy	
	Osteopathy	615.828
	Naturopathy	615.83
	Veterinary Science	619
	etc.,	

The result is substantial regrouping of chunks of the classification which will ultimately require a new system of notation. This notation may or may not be hierarchical in nature. While a decision on this has not been made, in general it is assumed that it will be "UDC-like" in nature and it is absolutely essential that it be independent of language.

Order of facets is based on Bliss which uses "standard citation order" as defined by the Classification Research Group, which is as follows:

(End Product)
Its Types
Its Parts
Its Materials
Its Properties
Its Processes
Operations on it
Agents of action
Place
Time

Of course not every topic will have all of the facets, and subjects will vary in the facets which they contain. The primary facets in Human Biology are:

TYPES of persons
PARTS, organs, systems of the human organism (e.g., nervous system)
PROCESSES of the human organism
Developmental processes (e.g. heredity)
Physiological processes (e.g. breathing)
COMMON facets (e.g. study, research, etc.)

In Bliss when this order of facets is applied at the highest level of Human Biology we find the following:

HUMAN BIOLOGY (General)
Principles, theory
Physiology & anatomy
(General properties)
(Physiology)
Biophysics (general)
Biochemistry
Metabolism
etc.
(Special physiological processes)
(By energy form and interaction)
Biomechanics...
etc.
(Processes special to given parts, organs, systems)

PROBLEMS TO BE ADDRESSED IN THE PROJECT

The process is far from straight forward. Some of the major problems have already been identified and are detailed in the McIlwaine and Williamson article (1993). Not all of them have been

PROCEEDINGS OF THE 5th ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP

resolved. The problems range from the general problem of the overall arrangement of the UDC Medical Sciences Class down to the details of levels of division in the hierarchies, length of arrays, location of fringe topics and the type of notation to be assigned.

Not unexpectedly the two classifications have been developed to different levels of division in different parts of the schedules. In many places BC2 is more detailed than UDC, but it is not possible simply to make an arbitrary decision that at a certain point a cutoff point occurs. This must be done in relation to the demands of the literature and the needs of the specialist. Therefore each term has to be considered on its merits and when the whole exercise has been completed, it will be necessary to examine the discarded vocabulary to see whether terms have been omitted that should not have been. Also terms included need to be measured against actual literature to ensure that none is superfluous, hence the reason for making use of the MEDLINE database.

During the pilot project there was some experimentation with levels of detail in BC2 and findings appear to indicate that second and third levels of BC2 may be appropriate. On this basis a decision has been made to include all facets provided in BC2 and as a general policy to include subdivisions up to 2 or 3 levels below the facet level. However, the rich vocabulary of both databases should be preserved. Tentatively the terminology from the lower levels of the BC2 hierarchies are being upward posted as "including" notes in order to accommodate additional vocabulary within the UDC without the need to incorporate the actual terms into the hierarchies and notate them. The investigators recognize that the same level of division may not be appropriate in all cases and that in the initial stages of the project the tentative policy should serve as a benchmark to be tested. Exceptions may be required.

The difficulty of dealing with fringe topics is another major problem. It is quite easy to decide that a large section of "fire precautions" is not required in medicine and that it can be removed from UDC class 61. It has, however, to be relocated elsewhere in the classification and whenever one attempts to carry out a bottom-up revision of a discipline within the context of a general classification that is in daily use, there is a knock-on effect. Sections of knowledge cannot be left in limbo and other classes not scheduled for revision cannot suddenly accommodate a whole new section without considerable adjustment. Users must be considered. While one can give warning that a whole class is up for revision, it is much more difficult to scatter fragments of a class within other sections of the classification. It is a two way problem because there are sections of both BC2 and UDC that are not wanted in the position they presently occupy. A case in point in BC2 Class H is the management of hospitals much of the detail for which UDC would derive from the Management class through the use of the colon.

CONCLUSION

To this point by no means all of the problems have been addressed or even identified. The notation is a case in point. Moreover, since there is no precedent for this research, it is important to recognize that procedures and policies may change as the work proceeds. This is necessarily exploratory research and much depends on finding solutions to problems encountered as work progresses. At this point in time the signs are hopeful, but the outcome is uncertain. Hopefully, whether the results are positive or negative, the work will provide some contribution to a better understanding of the role for classification in knowledge organization for the future.

BIBLIOGRAPHY

- Aitchison, Jean (1986). "A Classification as a Source for a Thesaurus: the Bibliographic Classification of H.E. Bliss as a Source of Thesaurus Terms and Structure", *Journal of Documentation* 42 (September 1986): 160-181.
- BSI ROOT Thesaurus* (1988) 3rd ed. Milton Keynes, Eng.: British Standards Institution.
- Goedegebuure, Ben G. (1993). "The UDC Consortium". *Extensions and Corrections to the UDC, September 1993*. (Series 15) The Hague: UDC Consortium. pp. 7-10.
- Marosi, A. (1969) "EURATOM Thesaurus and UDC: Combined Use for the Organization of Small Information Service." *Journal of Documentation*, 25 (September 1969): 197-203.
- McIlwaine, I.C. *Guide to the Use of UDC: an Introductory Guide to the Use and Application of the Universal Decimal Classification*. With the participation of Andrew Buxton. The Hague: International Federation for Information and Documentation (FID), 1993.
- McIlwaine, I.C. and Williamson, N.J. (1993). "Future Revisions of the UDC: Progress Report on a Feasibility Study for Restructuring". *Extensions and Corrections to the UDC, September 1993*. (Series 15) The Hague: UDC Consortium. pp. 11-17.
- Reisthuis, G.A. and Bliedung, Steffi (1991). "Thesaurification of the UDC". *Tools for Knowledge Organization and the Human Interface*. Edited by Robert Fugmann. Frankfurt/Main: Indeks Verlag. v.2, pp. 109-117.
- Williamson, Nancy J. (1992) "Restructuring UDC: Problems and Possibilities". *Classification Research for Knowledge Representation and Organization: Proceedings of the 5th International Study Conference on Classification Research*. Amsterdam: Elsevier, 1992. pp. 381-288.

PROCEEDINGS OF THE 5th ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP