Paper based on the talk presented at the at the Librarian Workshop in conjunction with "The 31st Annual Conference of the German Classification Society on Data Analysis, Machine Learning, and Applications", March 7-9, 2007, Freiburg i. Br., Germany [not published in the proceedings]

Enhancement of UDC data for use and sharing in a networked environment

Aida Slavic and Maria Ines Cordeiro and Gerhard Riesthuis

UDC Editorial Team
UDC Consortium, PO Box 90407, 2509 LK The Hague, The Netherlands
mail@udcc.org

Abstract. The authors give an overview of the present context of the development and use of Universal Decimal Classification (UDC) and describe the current work towards enrichment and improvement of the UDC data. UDC is a classification system used worldwide for organization and information access in various kinds of collections and information services. From 1993 UDC has been available for distribution to publishers and users as a database file which has initially improved the classification's potential for online use. From 1993 to 2007 the UDC vocabulary has been extensively revised and 14 updated UDC files have been distributed, on an annual basis. In order to take full advantage of this electronic availability, new developments should be carried out to support a better alignment between the forms and formats in which the UDC scheme is distributed and the classification's practical use in networked information services. The authors discuss different aspects that may concur with this aim, at the data structure, data content and data transmission levels. Examples include the possible provision of: an enhanced UDC data format; UDC multilingual data and UDC mappings to other subject indexing systems; UDC data exports in various other data formats that can be easily implemented, updated, shared or exchanged. In this context, a summary of current initiatives of the UDC Consortium is provided, notably in terms of renovation of the UDC technological management.

1 Introduction

The Universal Decimal Classification (UDC) was created between 1896-1907 as a classification for the detailed indexing of a vast and all-inclusive bibliography by Paul Otlet and Henry Lafontaine. They used the Dewey Decimal Classification for the basis of the new classification scheme and developed upon it an entirely new system in terms of structure, notation, syntax and vocabulary which became the first fully analytico-synthetic classification. Up to 1991, UDC was owned and developed by the International Federation for Documentation (FID). During the period of FID ownership, various editions of UDC were translated into 39 languages and continued to be used in all countries that, at some point, participated in the then strong international FID activity (c.f. Slavic (2004)). The strength of the FID declined in the 1980s and it was finally decided, in 1992, that the ownership of UDC be transferred to an international body of publishers: The UDC Consortium (UDCC). The Consortium is a not-for-profit organization which maintains, develops and distributes UDC and financially sustains itself from its international membership fees, selling licenses of UDC data and some publishing activity. Today, according to a recent world-wide survey, the UDC is used at least in 124 countries. In 34 countries in Europe, Asia and Africa, it is the main classification system used across of national library and information networks (c.f. Slavic, (2007)).

Upon the establishment of The UDC Consortium, to make maintenance, revision and distribution feasible, it was decided to take, as a basis, only 60.000 UDC numbers (corresponding to what was usually known as the medium edition) from the existing total of more than 200.000 available in the FID full edition. UDC was then stored in a database which has become a regularly maintained and developed UDC standard, known as the UDC Master Reference File (UDC MRF). Since the creation of the UDC MRF in 1993 (using CDS/ISIS software) users and publishers can buy UDC data in the form of a database file (ISO 2709) or a database text export.

Reduction of the schedules to a manageable size and their 'automation' allowed UDC content to be intensively revised and improved and a new version of its 'standard edition' to be distributed to users each year. The revision under the management of the UDCC took on board some suggestions put forward in the 1970s and 1980s by several experts, the most prominent of whom was Ingetraut Dahlberg. The new revision team, with I.C. McIlwaine as the editor in chief and with expert help from the then UDC editor of BSI editions, Geoffrey Robinson, continued in 1993 to update the vocabulary and clean UDC of cross classification, enumeration of complex subjects and remaining cultural bias. In the process of updating and creating new classes, efforts were made to apply facet analysis in a consistent fashion with the guidance of and input from Vanda Broughton, a member of the British Classification Research Group and one of the editors of the Bliss Bibliographic Classification (c.f. McIlwaine and Williamson (1994)). These changes further improved the synthetic capability of the UDC and allowed for the creation of structurally sound classes of simple concepts that can be logically combined to produce complex and more detailed, coordinated subject statements.

In the period 1993-2007 there were in total 14 new editions of the UDC with the MRF database now containing over 67.000 records. An overview of the changes produced since 1993 can be found at http://www.udcc.org/major_changes.htm. The original plan to revise the whole classification in ten years proved to be unfeasible, thus the revision work is still ongoing and will continue at a similar pace in the years to come, especially regarding the less revised parts of the scheme: natural and applied sciences.

Creation of the UDC MRF enabled the production of various electronic editions and facilitated translations. For instance, UDC schedules were published on CD-ROM in Spanish, Czech/English, and Russian and UDC on the Web has been made available since 2001: in English by the British Standards Institution and in Czech by the Czech National Library, while excerpts from the MRF are available online in Italian, Swedish and Spanish. In addition, new printed editions with the full content of the MRF appeared in recent years in Spanish, Russian, Ukrainian, Hungarian, French, Portuguese, Serbian, Croatian and English.

The availability of schedules in an electronic form encouraged research into UDC and a variety of implementations of UDC in information retrieval during the period 1993-2007. Riesthuis (1998, 1999), for example, devised algorithms for the decomposition of synthesized UDC numbers and their automatic linking to words based on the MRF. In the same period, UDC was used for resource discovery and automatic classification of Internet resources (an overview is given in Slavic (2006)). Also, Frâncu (2003) explored using UDC in the creation of multilingual thesauri. More recently, Baliková (2005) reported UDC as being used in supporting cross national and cross language subject access and Schallier (2005) described methods for improving subject browsing in library OPACs.

2 Characteristics of UDC as an Indexing Language

At the time of its creation, UDC represented an innovation in the field of documentary classifications and inspired further developments which led to purely faceted systems. While the main UDC schedules with their disciplinary organization remained superficially very similar to Dewey, the deeper structure and rules for development and use of the vocabulary were designed in a different way. From the very beginning the UDC schedules were organized in such a way that generally applicable concepts (common auxiliary tables) are kept as separate facets so they can be re-used and combined across disciplines and fields of knowledge with any subject within the main discipline hierarchies. In addition, each main discipline/class can have discipline-specific concepts (processes, operations, agents, materials etc.) aggregated into facets called *special auxiliary tables* and these can be combined with any main class number within the respective discipline. But most importantly, provisions were made for every class number in UDC to be combined with any other class to express complex or interdisciplinary or multidisciplinary subjects (Fig. 1).

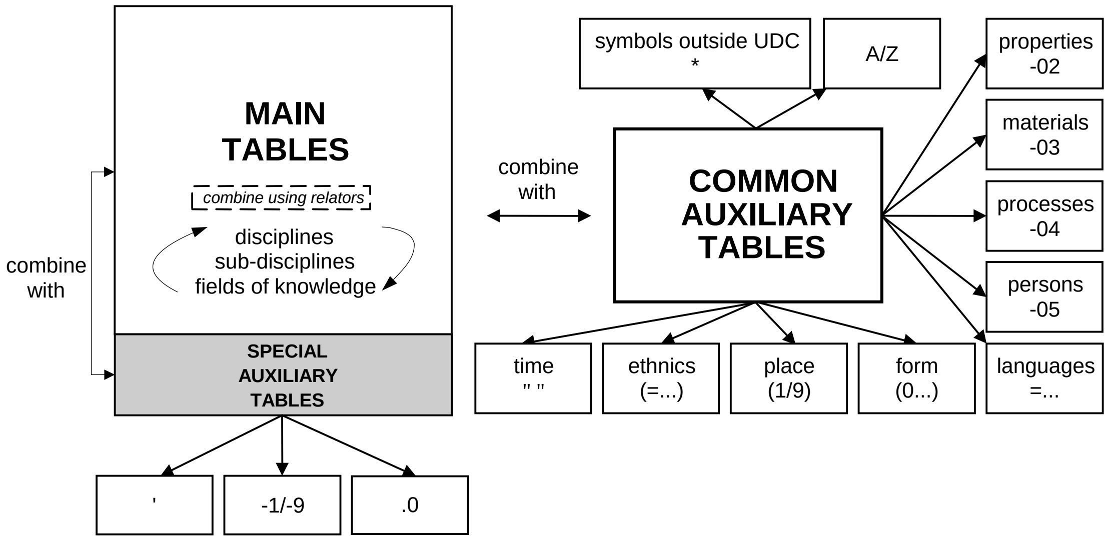

Fig. 1. UDC tables

The synthesis is facilitated by: a) notation devices known as facet indicators (punctuation marks) and b) relational symbols. Facet indicators (or their absence) denote the type of table from which a number is taken and relationship symbols denote the type of relationship between two numbers. Thus simple numbers are clearly delimited one from another allowing for easy composition and decomposition of synthesised UDC statements. For example:

(046)	Newspaper article	common auxiliary table (Form)
(430)	Germany	common auxiliary table (Place)
(430***-22***)	Germany - rural	common auxiliary table + ***sp.aux table***
338.48	Tourism	main table
338.48***-6***	Tourism - according to motive	main table + ***sp.aux table***
502.17	Protection of the environment	main table
:	relationship symbol	

Combinations:

338.48-6:502.17	Eco-tourism
[338.48-6:502.17](430-22)	Eco-tourism in rural Germany
502.17(430)	Protection of the environment in Germany
502.17(430)(046)	Newspaper article on the protection of the environment in Germany

Therefore, UDC vocabulary does not have to be large and yet it can be used to generate a very large number of subject representations, at a very detailed level. More importantly, indexers can 'build' numbers for new and emerging subjects or for the innovative treatment of existing phenomena. In addition, they can determine their own method of collocating subjects by changing the usual citation order of the elements in the synthesised numbers, thus using classification to support the particular needs of their collection as suggested by Robinson (2003). But because UDC can also be used in a simple way as a disciplinary, hierarchical classification, it is widely adopted for shelf-arrangement of books and non-book material. In fact, libraries often combine shorter and simple UDC numbers for shelf arrangement and more detailed, synthesized, UDC numbers for content indexing and information retrieval. Depending on the size and purpose of a collection, a local indexing policy may determine the level of complexity and specificity: simple UDC numbers or synthesised UDC numbers (c.f. Slavic, (2004a)).

3 Some Aspects Concerning the Automation of UDC

When using UDC in an information system, one should make the most of its hierarchical and synthetic structure without compromising the accuracy of indexing, cost efficiency of the system and user friendliness of the interface. All this can be achieved if the classification vocabulary is managed through an authority control function, that is, having classification data managed as an independent component of the system that 'feeds' into both metadata population and information retrieval.

An authority control function, and subsystem, addresses the problem of reusing the same authorised entries, thus saving the cataloguer/indexer's time and ensuring consistency. This means, for instance, that once a synthesized UDC number is described as an authority entry, it can be linked to as many metadata instances as necessary without the need for retyping or synthesising the same elements again. Other advantages of this approach are the separation of data content from its presentation at the interface level and the creation of multiple access points to any given vocabulary entry. As a result, classification numbers can be replaced by words at the user interface while the semantics of a classification structure can be fully exploited in browsing or searching. Especially advantageous in the case of UDC is that in an authority entry, a synthesised classification notation can be managed in a such a way as to allow search and access to every meaningful element of the notation; if words are linked to the components of a notation, classification retrieval using words, in one or more languages, is feasible. Furthermore, in an authority file UDC can be linked to other indexing languages such as subject headings or thesauri or other classifications. In addition, a classification authority file allows automated updates and global changes of data, and unlimited linking and coordination of data elements. Finally, classification authority records can also store other data elements essential for subject browsing and search expansion, such as links between any UDC number and its broader and related classes, which are not always explicitly declared in the classification scheme itself.

It is safe to assume that managing subject access via authority control will gradually become the predominant model in all information retrieval systems that use classifications. This approach is especially relevant for metadata based information retrieval across distributed, heterogeneous collections, for which authority control may significantly improve subject access while preserving idiosyncratic indexing practices of individual collections. In the bibliographic domain, the traditional practice of authority control of subject access points has not been addressed with respect to classifications when compared with alphabetical systems. In fact, general purpose MARC authority formats (e.g., *UNIMARC Authorities*) do not offer sufficient provisions for the requirements of managing classification access points. But special authority standard formats for classification authority files have already been created and will continue to be improved (cf. *MARC 21 Concise Format for Classification Data* (2005), Concise *UNIMARC Classification Format* (2000)).

When the UDC MRF database was created its main objective was to support the maintenance and updating of content and to make it easier for translations and the preparation of printed editions. Nowadays, UDC publishers normally import MRF in their own databases to manage translations, publishing processes or the preparation of electronic editions. Some users (e.g., libraries) use MRF to feed their authority files or systems of a similar function. The advantage of loading UDC data upfront, to make the indexing language available online to the cataloguer, is still in need of exports more suited to that purpose.

In all cases where the MRF is used, each user has to build and maintain local tools to process MRF data according to their needs and purposes, usually to integrate MRF data into their particular systems. The effort/cost implied in these integration activities is related to the ‘readability’ and usability of data by non-expert (systems and people) in both CDS/ISIS and UDC. Semantically and syntactically, UDC itself is complex enough to make it a specialist system. The existing export formats created in 1993 either reduce complexity by almost completely removing the data structure (as with current ASCII exports) or, to use the data structure, they imply additional specialist knowledge (namely of ISO 2709), which is not mainstream. ISO 2709 knowledge is usually available in library systems but not found so easily elsewhere. Besides, even in the case of users having knowledge of this standard, the syntax/semantics of MRF have nothing in common with MARC formats, thus always requiring the definition and maintenance of mappings and conversion programs, an effort repeated wherever that is needed. One way to alleviate these inconveniences is to provide UNIMARC and MARC21 exports of UDC data, according to the MARC21 and UNIMARC Classification Formats.

Another trend to facilitate the use of specialist data by different information and technological communities is to have XML expressions of it, thus opening such data to processing by a variety of XML tools. This recommends the existence of XML exports of UDC data, from MRF, UNIMARC and MARC21 formats. This strategy should be followed by possible exports according to SKOS (XML/RDF) and XML Topic Maps, at a later stage (c.f. SKOS Core Guide (2005); ISO/IEC 13250 (2003-2007); XML Topic Maps (2001)).

4. A New UDC Editorial System

The editorial activity of revising UDC schedules and the management of the MRF database are closely linked and they are both connected to the strategy for developing the UDC products. All aspects discussed in the previous sections have been taken into consideration for the requirements of a new UDC editorial support system, which has been under development since the beginning of 2007, aimed at improving UDC content management, revision process and product management (c.f. Cordeiro and Riesthuis (2006)).

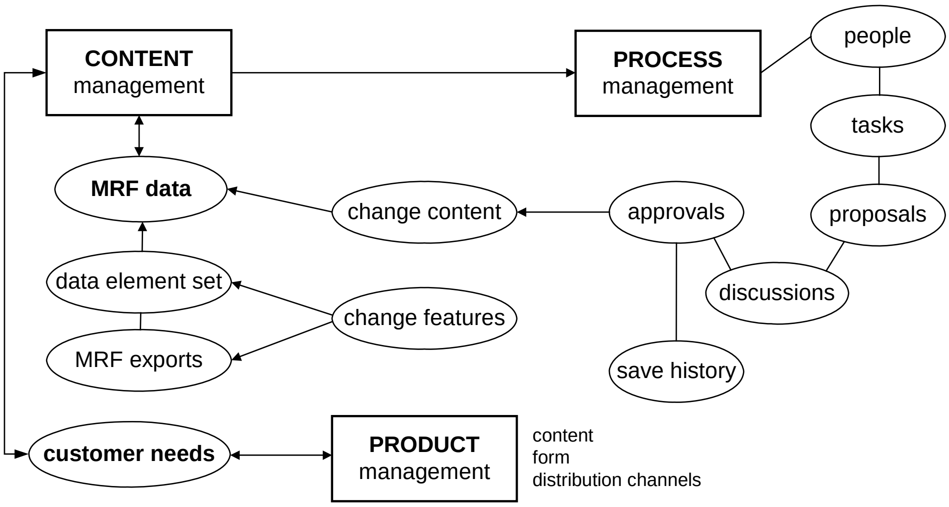

Fig. 2. The rationale for the new UDC editorial support system

The main objectives of the new system are: i) to ensure sustainability of the UDC MRF through a more powerful and flexible database system and to improve system security and portability; ii) provide a online collaborative system for UDC revision which will facilitate inclusion and management of a larger team of internationally distributed editors; iii) to make possible management, tracking and archiving changes; and iv) to diversify MRF products: expand to MARC and ontology formats and their XML derivatives and include provisions for future inclusion / export of multilingual data and data regarding mapping to other subject systems.

5 Improving Data and Diversifying Exports

The migration of the UDC MRF to a new database provides an opportunity to make a number of improvements in both the MRF data and exports.

5.1 Improving of UDC MRF format

a) Changes to existing fields:

- **UDC heading field** (notation) - the existing field that now allows for UDC notation to be entered as a simple text string will be changed to provide coding for each element of a synthesised notation in order to enable computer-readability of the individual components of a complex UDC notation.
- **Language codes**. All textual fields of the MRF that now allow just a limited number of languages will get provisions for a full set of standard-based language codes.

b) Introduction of new elements

- **Unique Identifier** - the current database has no unique identifier for each MRF record; this will be added to allow maintenance of database consistency in between versions, exports, etc.; and it will also facilitate automatic updates. This means that every class will get a notation-independent unique and permanent identifier which is essential for the MRF to function as a true UDC registry.
- **Hierarchical next higher notation** - although UDC notations appear to be expressive - this being true in the majority of cases - there are many exceptions. It is a well known fact that all classification systems with decimal notation often have 'jumps' in their hierarchy, 'false' hierarchies or hierarchies that span several coordinated classes. Data on broader class for each notation is needed in order to automatically and correctly build the hierarchical chains. Once the field is implemented, linking content can be generated automatically, followed by manual inspection and correction of the exceptions.
- **Notation history** - this element will allow linking of the revised classes to the old notation which will be useful to both the UDC editorial team and users in order to trace and manage changes.
- **Link of each notation to its special auxiliary tables -** currently, MRF records do not specify for each notation the range of special auxiliaries that can be combined with it; normally, special auxiliaries are listed at the beginning of the class only and are applicable to all members of the class if the opposite is not stated. These kind of 'implicit' rules, or those explained in the textual instructions, have to be made machine-readable throughout the class in order to facilitate management and exploitation of UDC.
- **Index entries (keywords)** - this field will hold uncontrolled verbal expressions relevant for the retrieval of the record; they can also be used for preparing subject alphabetical indexes of printed editions.
- **Mapping** - provision will be made to store mapping to other subject representation schemes as part of future developments.

5.2 Corrections and consistency checking of MRF

Modifications of the data structure, as mentioned above, will require a great deal of data processing for the changes to take actual effect with the current content of MRF. Migration of data will also create an opportunity to validate, clean-up and correct mistakes regarding data consistency that we are already aware of.

5.3 New content

Multilingual schedules. Multilingual editions have been discussed for some time among UDC Consortium members. Users have also expressed interest in multilingual schedules, especially for electronic and online applications. The old database system, however, had limitations with respect to supporting different languages and scripts and all projects with respect to this had to be delayed. The requirements for the new editorial system in preparation took these needs into consideration and the future MRF database will have the conditions adequate for a multilingual UDC data repository.

Mappings. In practice, UDC is often used in combination with other subject indexing languages. It is therefore envisaged, and considered advantageous, to gradually enrich the UDC MRF with mappings to other systems. While the mapping itself or its use may be, in some cases, subject to a copyright agreement - an aspect that has to be part of future projects, managed separately - it is useful to preserve such valuable data, whenever available, in the MRF database. Mappings may be useful whether available across all subjects to a certain level or in a selection of subjects. They may include links to some special classifications such as the *National Library of Medicine (NLM) Classification*, the *Subject Mathematical Classification* (AMS SMC) or the Physics & Astronomy Classification Scheme (PACS); or to some general national classification systems such as the *Nederlandse Basisclassificatie*, the *Sveriges Allmänna Biblioteksförening* or some internationally used schemes such as the *Bliss Bibliographic* or *Dewey Decimal* classifications. The same is valid for subject heading systems such as the *Medical Subject Headings* (MeSH) or thesauri such as the *AAT - Art and Architecture Thesaurus*.

5.4 Improving UDC exports

In addition to the objectives and opportunities of the new editorial system, we have also introduced the possibility to redefine and improve MRF exports. As mentioned before, the new system should provide more export services than the currently provided MRF ISO 2709 and ASCII exports. Additional exports envisaged at this point include an ISO 2709 UNIMARC export (to be defined according to the UNIMARC Classification Format), an XML version of the current MRF export and exports for XML expressions of the data in UNIMARC and MARC21, in conformance with MARCXchange. The definition of these exports will require an analysis of the mappings between MRF and MARC formats, thus being again an opportunity to detect aspects of UDC data format that can be improved. The new exports will increase flexibility for current users and will possibly make UDC more attractive for new implementers more orientated towards classification use in a networked environment. It should be emphasised, however, that the maintenance of the current exports will ensure continuity for those wishing to maintain the tools and procedures they already have to deal with UDC data as it has been provided so far.

6 Conclusion

The emergence of online usages of classification systems combined with the requirements to access, share and search classification data in a networked environment have raised new demands with respect to the form and format of UDC data availability. Towards the end of the 1990s, both publishers and users have expressed their need for having UDC in formats that can make publishing and implementation easier, quicker and cheaper, and also to be able to do automatic updates and use the scheme with multilingual features. These issues have now started being addressed, and have good prospects with the new UDC editorial system, under development since 2007.

The present initiative of the UDCC to renew the technological infrastructure of the UDC Master Reference File corresponds to a strategy wider than just the modernization of the database. This strategy encompasses the renovation of the very processes of collaboration upon which the evolution of the UDC scheme depends, as well as the processes by which UDC is distributed to publishers and other users of UDC machine-readable files.

References

BALIKOVÁ, M. (2005): Multilingual subject access to catalogues of national libraries (MSAC) : Czech Republic's collaboration with Slovakia, Slovenia, Croatia, Macedonia, Lithuania and Latvia, 71th IFLA General Conference and Council, August 14th - 18th 2005, Oslo, Norway. http://www.ifla.org/IV/ifla71/papers/044e-Balikova.pdf

Concise UNIMARC Classification Format (2000).
Available at: http://www.ifla.org/VI/3/p1996-1/concise.htm

CORDEIRO, M.I. and RIESTHUIS, G. J. A. (2006): A new editorial support system for UDC. Extensions & Corrections to the UDC, 28, 17-22.

FRÂNCU, V. (2003): Multilingual access to information using an intermediate language : proefschrift voorgelegd tot het behalen van de graad van doctor in de Taal- en Letterkunde aan de Universiteit Antwerpen. Antwerpen
http://www.bcub.ro/continut/noutati/v_francu_doctoral_thesis.pdf

ISO/IEC 13250 (2003): Information technology. SGML applications. Topic maps. Geneva: ISO.

ISO/IEC 13250-2 (2006): Information technology. Topic maps. Part 2: Data model. Geneva: ISO.

ISO/IEC 13250-3 (2007): Information technology. Topic maps. Part 3: XML syntax. Geneva: ISO.

McILWAINE, I. C. and WILLIAMSON, N. J. (1994): A feasibility study on the restructuring of the Universal Decimal Classification into a fully-faceted classification system. In: H. Albrechtsen, S. Oernager (Eds.): Knowledge organization and quality management : proceedings of the Third International ISKO Conference, 20-24 June 1994, Copenhagen, Denmark. Indeks Verlag, Frankfurt, 406-413.

MARC 21 Concise Format for Classification Data (2005): Update No. 6. Library of Congress. Available at: http://www.loc.gov/marc/classification/eccdhome.html.

RIESTHUIS, G. J. A. (1998): Decomposition of UDC-numbers and the text of the UDC Master Reference File. In: W. Mustafa Elhadi, J. Maniez, S. Pollitt (Eds.): Structures and relations in knowledge organization : proceedings of the Fifth International ISKO Conference, Lille, 25-29 August 1998. Ergon, Würzburg, 221-228.

RIESTHUIS, G. J. A. (1999): Searching with words : re-use of subject indexing. Extensions & Corrections to the UDC, 21, 24-32.

ROBINSON, G. (2003): Citation order in the UDC. Extensions & Corrections to the UDC, 25, 19-27.

SCHALLIER, W. (2005): Subject retrieval in OPAC: a study of three interfaces. In: Jesús Gascón, Ferran Burguillos and Amadeu Pons (Eds.): La dimensió humana de l'organització del coneixement, 7 congreso del Capítulo Español de ISKO, Barcelona, 6-8 de julio de 2005. Universitat de Barcelona Departament de Biblioteconomia i Documentació, Barcelona, 557-567. Also available at: http://dlist.sir.arizona.edu/1323/

SKOS Core Guide (2005): W3C Working Draft 2 November. Alistair Miles and Dan Brickley (Eds). Available at: http://www.w3.org/TR/swbp-skos-core-guide/.

SLAVIC, A. (2004): UDC translations : a 2004 survey report and bibliography. Extensions & Corrections to the UDC, 26, 58-80. Available at: http://dlist.sir.arizona.edu/649/.

SLAVIC, A. (2004a): UDC implementation : from library shelves to a structured indexing language. International Cataloguing and Bibliographic Control, 33 (3), 60-65. http://www.ifla.org/IV/ifla69/papers/032e-Slavic.pdf.

SLAVIC, A. (2007): Use of the Universal Decimal Classification: a worldwide survey. Journal of Documentation, [in print]. Pre-print available at: http://dlist.sir.arizona.edu/1555/

XML Topic Maps (XTM) 1.0 (2001): Steve Pepper and Graham Moore (Eds). Available at: http://www.topicmaps.org/xtm/1.0/index.html.