Wouter Schallier
K.U.Leuven University Library, Belgium

What a Subject Search Interface Can Do[1]

Abstract: K.U.Leuven University Library (Belgium) developed an experimental interface for subject search by UDC in the OPAC. The interface combines the search facilities of a classification with those of a word system, since it enables the end user to search by subject terms and to see these terms in the hierarchy of broader, parallel and more specific terms. This project should be seen as an important indication of the library's growing concern to present its information sources in a content-structured and user-friendly way. At the same time, it has to be situated in a new policy for knowledge organization, which aims to find a balance between the local and overall needs of a library network. Finally, this project comes at a moment when K.U.Leuven University Library is in full conversion to Aleph 500 software.

1. Local Versus Overall Needs

Knowledge organization is a complicated matter in a networked world. Our networks make an enormous amount of information easily available. However, this information is only accessible if it is well organized. Network rules and standards have to be established and respected to guarantee information consistency and to facilitate information exchange within and outside the network. K.U.Leuven University Library[2] is part of LIBIS-net, the largest network of academic libraries in Belgium; it is also a network by itself, since it has a decentralized structure consisting of several department libraries and one central library. The use of standards and overall network rules results in consistent information and allows partner libraries and their end users to profit fully from the advantages of a network. Unfortunately, standards and overall solutions have more than just advantages. They also have limitations, since they do not necessarily give suitable answers to the needs of a specific library. One cannot easily convince a department library to use a universal and generally accepted thesaurus like Library of Congress Subject Headings (LCSH), if the users have grown familiar with a specific and local classification that is much better adapted to the collection.

Since 1977, K.U.Leuven University Library has used the DOBIS/LIBIS library system, which offers a technical solution for both overall and local needs. The main part of the bibliographic description is stored on the overall level and is visible for the whole network. Universal Decimal Classification (UDC), Library of Congress Subject Headings (LCSH) and Medical Subject Headings (MeSH) are all generally accepted subject-cataloguing methods, and therefore they are used on the overall level. Local information, such as thesauri and classifications used exclusively in a specific library, are stored in local files ("local keywords") and can only be seen by and in that library (up to branch level).

2. Elements that Hinder Good Subject Access

Statistics show that in our catalogue 27.04% of all documents and 38.16% of those younger than 1992 have "local keywords"[3]. However, these percentages are very misleading. Many "local keywords" have nothing to do with subject cataloguing and are in fact purely administrative information. The percentage of real subject information on the local level is difficult to determine, since no formal distinction is made between administrative and real subject information.

On the overall level, 20.73% of the documents in our catalogue have at least one (manually given) UDC code, and the relative use of UDC has remained stable in recent years (21% of the documents younger than 1992). 12.58% of all documents have (automatically uploaded) LCSH, and if we consider only the documents younger than 1992 we arrive at a percentage of 20.12. Only 1.81% have MeSH, used by our Library of Biomedical Sciences [4]. Apparently, UDC has been the most important and most stable tool for (manual) subject cataloguing, even if in recent years (automatically uploaded) LCSH has gained ground. Nevertheless, the percentages just mentioned are very disappointing. A higher rate of documents catalogued by subject is absolutely needed to provide representative search results. The only available tool for manual subject cataloguing on the overall level, UDC, is used by a dedicated, yet small, part of our librarians. The others do not use UDC systematically, replaced it by their own local classification or thesaurus, or even do not index or classify their collections by subject at all.

A second problem lies in the local files that can only be viewed on the site of the department library to which these files belong. Thus a user, who consults the OPAC at home or even in another department library, is unable to take advantage of the local subject information. This makes the potential use of the local thesauri and classifications very limited, and reduces them in practice to a professional tool for the librarian. The lack of user friendliness is compounded by the large vocabulary and syntactical variety of the local indexing practice and the absence of indexing guidelines. User training which is regularly organized in most department libraries can partly compensate this.

Another difficulty is the way the subject search option is presented in our OPAC. Few, if any, users know that searching by “classification” means searching by UDC or that this is a method to search by subject. Indeed, the OPAC contains a so-called UDC authority file with the UDC codes and the descriptors in three languages (Dutch, English and French). Both UDC codes and descriptors are entirely searchable. They are linked to the UDC codes in the bibliographic descriptions, which allows us to retrieve documents about a specific subject. The potential of this tool is high, but in reality UDC is, just like the local subject information, mostly a professional tool for those librarians who use it for their own collections. Only few beyond the library staff know that for specific subjects searching by “classification” gives good results. On the other hand, searching by LCSH is also problematic, since these subject headings are in English, and many users are not aware of that (not many read our search tips) or do not feel comfortable in using them.

3. A New Breath for Subject Cataloguing

All these elements together make consistent knowledge organization a bit problematic. This does not mean there is no need for good knowledge organization, on the contrary. All our librarians feel the need to organize their collections by (broad) subject. Those who do not use a local thesaurus or classification, at least organize their collections physically, often by thematic shelf numbers. Of course, this is useful but our end user rightly asks more than physical or local organization of the collections. A survey made in 2002 on the academic staff’s needs for e-services tells us that 75% of the respondents rank “more keywords for the whole library collection” in the top 5 of priorities[5].

How can we give an appropriate answer to this demand? Despite all technology, good subject cataloguing remains a time-consuming task. If we want our librarians to invest in it, we

should be able to offer subject-cataloguing methods on the overall level that are flexible enough to be adapted to librarians' local needs.

There was no dispute about the use of MeSH, since these subject headings are generally supported in Biomedical Sciences. Nor was the continued use of LCSH doubted, these being an automatic and regular enrichment of our catalogue, without intervention of any indexer. We even considered enriching our bibliographic descriptions with Dewey Decimal Classification in the future, on the condition that this would happen in the same automatic process as for LCSH.

But even more important was that we agreed that, for us, UDC remains the best basis for manual subject cataloguing on the overall level. UDC is a generally accepted universal classification, is very flexible and has great potential as a faceted system. It is hard for us to find a better solution than the multilingual UDC authority file we developed for the needs of our network.

A more flexible use of UDC should be encouraged. Most important would be to update the UDC codes and their descriptors to make the UDC authority file a user-friendly tool. Changes or extensions in our authority file would be adopted more efficiently on the condition that they would be fully formalized. Some libraries already worked in this way in the past. Our library of Theology, for example, added several extensions to the official UDC, which allowed it to develop an elaborate theological classification and vocabulary based on UDC.[6]

UDC would allow us to build *virtual library shelves*, where a document's subjects are expressed in thematic categories rather than in detailed verbal terms.[7] Classifying documents, in the sense of grouping them in clusters, subject categories, requires less effort than fully indexing them into detail. In a modern OPAC context, where subject searches should involve an "intelligent" combination of subject and other relevant bibliographic fields (title words, author, etc.) as well as, and increasingly, abstracts and even entire documents, a classification scheme may well offer an adequate and quick answer to the demand for more efficient subject-search facilities. It is our experience that most end users are not familiar with large controlled vocabularies. Making a search in our catalogue should therefore be very instinctive and the hierarchical structure of UDC could be used to guide the end users through more or less broad subject categories. End users are familiar with internet search engines like Google but many of them also know the obvious advantages of web directories.

Finally, UDC would help us to achieve our goal of more consistent and more user-friendly knowledge organization. By linking shelf numbers and local subject cataloguing methods to UDC, we would be able to reflect the subject organization of the collections, which is often purely local or physical now, in a virtual organization on the overall level. Subject information would then become more visible and useful to the end user.

4. New Subject-Search Interface for E-sources

Once these principles had been set, it was a major task to achieve these objectives in practice. Several new projects were implemented. A new interface for the electronic sources (databases, e-journals, websites), called LibriSource, was created with special attention for subject search[8]. Each librarian was asked to draft a four-level hierarchical subject list for his area. The librarians were strongly advised to take into account generally accepted thesauri and classifications in their domain, if available. Their subject terms were linked in the background

to the appropriate UDC codes, which are not visible in the users' interface. In the administrative module the UDC codes are still shown, but the librarian does not need to be familiar with UDC since he can simply select a term from the subject list.

Asking our librarians to draft their own lists rather than compelling them to use UDC avoided endless discussions about whether UDC is adequate or not. In seeing their subject lists integrated in the overall list, the librarians discovered in practice the need for consistency in the structure and wording of the subject terms, not only on the local level (their own library) but especially on the overall level (the network and broader). They saw interesting applications of subject cataloguing, such as the generation of dynamic lists of e-sources organized by subject (for their web pages). At the same time, they wound up classifying the whole University Library's e-collection (some 12 000 titles) by UDC without feeling bothered by it.

5. New Subject Search Interface for the OPAC

To stimulate the use of UDC, it was of capital importance to improve the subject-search interface in the OPAC. The main idea was that the end user would be able to search by subject words and should not be confronted with the UDC codes. On the other hand, the UDC structure could be used in the background to show each term in its hierarchy of broader, parallel and narrower terms. In short, we would try to take maximal advantage of the potential of our UDC authority file, in which words and classification codes are combined in a unique way.

Before we discuss the UDC interface we developed, we should first describe very briefly the structure and characteristics of our UDC authority file. Our UDC authority file is based on different versions, namely the Dutch short edition, the English medium edition, the French full edition (class 2) and the version of Erasmus University Rotterdam. The use of auxiliary numbers and combinations is kept very limited. Only auxiliary numbers of language (=…), form (<0…>), place (<1/9>), ethnic group (<=…>) and time ("…") are allowed. Note that we replaced, for technical reasons, the ( ) of the official UDC by < >. All main numbers and combinations are explicitly or implicitly (by syntactical indications) present in the authority file. Main and auxiliary numbers are separated by a space to enable permutation. Some other classifications are also integrated in our UDC and are separated from UDC by *. 681.3* for example (which stands, in our authority file, for computer sciences), is followed by the ACM-classification.

In conjunction with LIBIS, our IT-department[9], we developed an interface that can handle a search by UDC words as well as by codes, all taken from our UDC authority file. Both words and codes can be truncated and can be combined with Boolean operators. Thus when we enter the search statement "Leuven?", we find 100 relevant UDC codes (fig. 1). It is possible to refine the results by selecting one of the 10 main UDC classes (fig. 2, where we selected class 2, theology) or by combining with another search term.

Fig. 2 also shows that efforts have been made to translate the Dutch descriptors (preceded by n-) into French (preceded by f-). Classes 5 en 6 (exact and applied sciences) are entirely in English. Although the presentation still needs work, it already reveals the multilingual potential of our authority file.

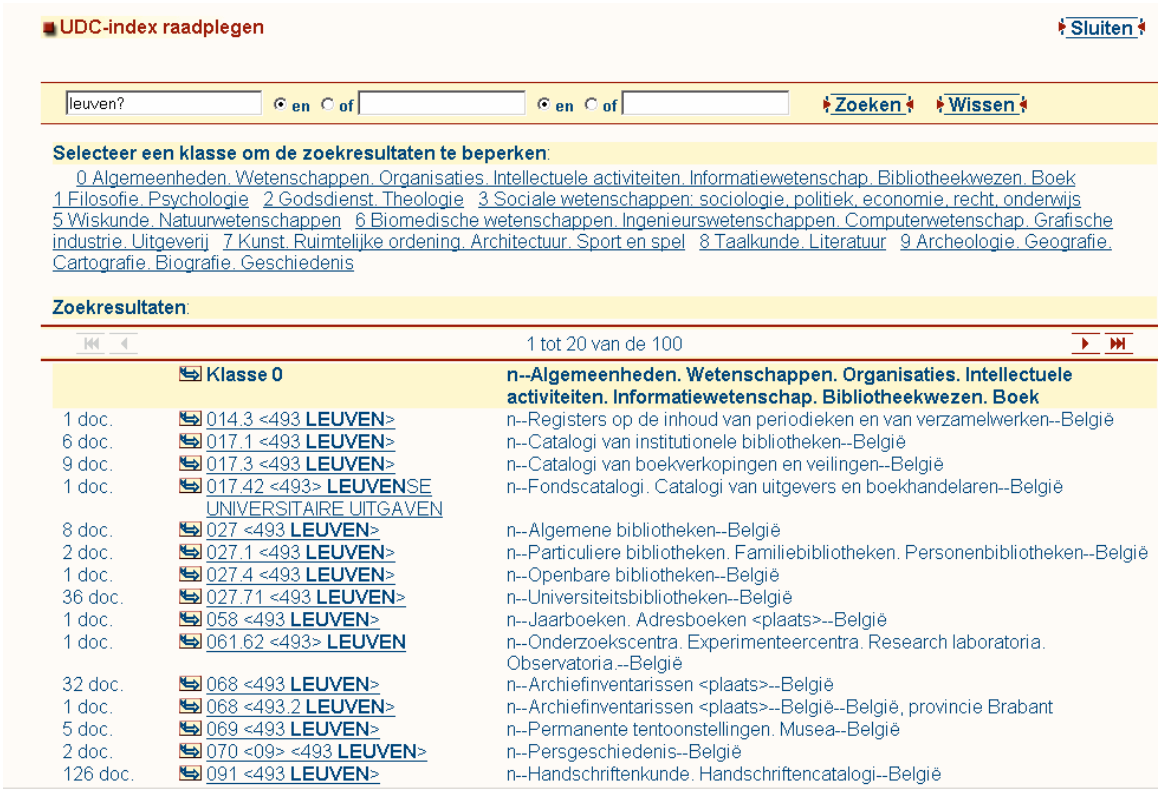

fig. 1 Search results for “Leuven?”

fig. 2 Refining the search result (fig. 1) by selecting UDC class 2 (Theology).

With the search statement “Leuven? AND jezuiet?” (fig. 3), we reach one result, namely 271.5 <493 LEUVEN> n--Jezuïeten--België. The broader search “Jezuiet? AND België” (or:

“Jezuiet? AND ?493?”) will find everything about the Jesuit order in Belgium, including but not limited to Leuven (fig. 4).

fig. 3 Refining the search result (fig. 1) by combining with the term “jezuiet?”

To the extreme left of each UDC code, we see the number of documents in the catalogue that have that particular code. They will be shown by clicking on the underlined UDC code.

fig. 4 Search results for “Jezuiet?” AND “belgie”

Clicking on the arrow at the left of each UDC code will show the placement of the code in its hierarchy. In fig. 5, we first made a search on 271.12 (Cistercian order). Then we clicked on the left arrow of this code, which provides a quick and structured overview of the subject (different aspects of the Cistercian Order in Belgium, the Cistercian Order in other countries, etc.) and allows us to refine (or broaden) the search statement without having to be familiar

with the vocabulary and the syntax. We see the first 20 of the 32 combinations with this code that are positioned one level lower in the tree structure. The broader terms are also clearly visible and the parallel terms of 217.12 can be found if we click on the left arrow of the broader term that is exactly one level higher (271.1).

Showing the semantic context of a subject term offers a solution for ambiguity, which is usually a problem for all kinds of controlled vocabularies. If a term is seen in its hierarchical context, the appropriate meaning will be easily identified. The hierarchical relations between the terms are laid by the algorithm (right truncation leads to the broader category) and not manually. This supposes of course that the structure of the UDC notation reflects by itself the right semantic relations. In most cases this is indeed true, but there are some well known examples like 563.4, where right truncation does not lead to its broader term (not 563 but 562 is semantically the broader term). [10] Algorithms will probably always fail here, as does ours.

Some UDC codes in our interface do not have documents, because they are purely syntactical indications. 271.12 <4/9 A/Z> (fig. 5) indicates that combinations with UDC auxiliary numbers denoting country (4/9) and city (A/Z) are possible. Thus we find 271.12 <41> (Cistercians in United Kingdom), 271.12 <44> (Cistercians in France), etc. The descriptors of these country codes are provided in the column to the right. The name of a city appears in the UDC code in the left column (e.g. 271.125 <493 ORVAL> in fig. 1). This is correct because we find the syntactical indication 271.125 <4/9 A/Z>. In a further development, the text in the UDC codes could be copied into the descriptors (right column), which would enable us to leave the UDC codes in the background (without loosing their hierarchical structure). Of course this would force us to visualize even more than now.

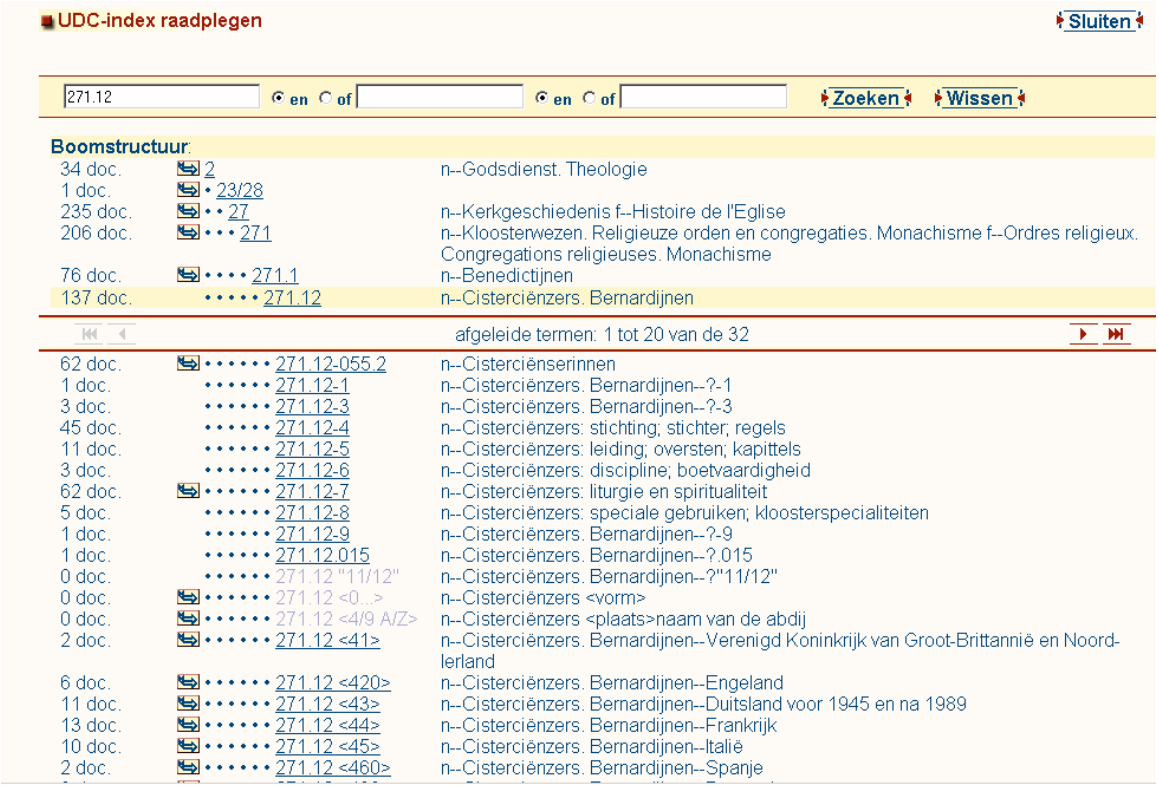

fig. 5 UDC 271.12 is shown in its hierarchical position

The left column of the UDC codes is in fact a combination of the index of UDC codes in the catalogue and the authority file. When the algorithm does not find a particular UDC code as such in the authority file (which is often the case with combinations of main and auxiliary numbers), it decomposes the code, e.g. 271.12 <45>, as being 271.12 in combination with <45>. Then it identifies the meanings of the two components in the authority file. As last step, it joins them again in one descriptor: 'n--Trappisten. Hervormde Cisterciënzers--Italië' (Cistercians in Italy). Each combination which is not found as such and is composed by the algorithm, can be recognized by a -- between the components. If the algorithm does not locate a meaning for a UDC code in the authority file, it puts a '?'. In fig. 5, we see the example of 271.12 "11/12". The algorithm correctly interprets the main number 271.12 as being "Cisterciënzers. Bernardijnen" but does not recognize "11/12", since this has not yet been defined in the authority file as being "$12^{th}/13^{th}$ century". Once "11/12" and its descriptor have been added to the authority file, all combinations with this auxiliary number will be interpreted correctly.

6. Further projects

In this interface for subject search by UDC, we tried to combine the advantages of a word system with those of a classification. A search by words is a much more natural approach for the end user than a search by UDC codes which are meaningless to him. On the other hand, the structure of the UDC permits us to situate each subject term in its semantic context. This aspect is often neglected in OPACs. Usually it is very hard to navigate through the structure of relevant broader, parallel and narrower terms. However, it is our experience that a clear overview of a subject domain would help the end user in finding information, since only a few of them are familiar enough with controlled vocabularies and their syntaxes.

The idea of combining a classification and a word system is not new of course. Behind every heading in the MeSH thesaurus there is a classification code, which permits a structured display of the headings, a facility that in most OPACs has not been exploited. In our OPAC we did use these MeSH classification codes in the background to construct the hierarchical relations. The result is that we can now search our catalogue by MeSH headings in the usual way and that we can also situate each heading in its hierarchy, including references, scope notes etc.[11]

The UDC and MeSH search interfaces went on line in our OPAC in September 2003. Though they are still in an experimental stage, we expect the UDC interface, in particular, to give a new stimulus to knowledge organization in K.U.Leuven University Library. Internal updates and extensions to our UDC authority file are constantly being made to respond better to the needs of our end users. The UDC authority file will be further translated into English, Dutch and French.

We already mentioned our use of UDC to build virtual library shelves. We put this in practice by correlating shelf numbers to UDC codes. The shelf numbers used in our libraries often group documents about the same broad subject. These clusters can be linked to a UDC code that allows us to give a UDC to these documents in an automatic process. Even if these UDC codes usually correspond to relatively broad subjects, this can be seen as a successful first step in classifying documents by subject. The advantages are important enough: most of the documents did not have any subject information at all before this operation and therefore could not be retrieved by a subject search. After this operation, they are classed in at least a broad subject category and can be retrieved by a UDC subject search in the OPAC. This is

clearly another approach than typical subject indexing, where often unique and very specific combinations of subject headings are given to documents. Classifying information is grouping documents about a similar subject in such a way that the end user can use each subject class as a pool of relevant documents. This pool can lose its relevancy when the number of documents is too high. Then the subject class should be split up into subdivisions.

Some people will certainly argue that (our way of) classifying does not give specific enough subject information about the documents. This could be right, were it not that subject search in modern OPACs is, or should be, an "intelligent" combination of subject terms, classification codes, title words, names, abstracts, publisher names, etc. Limiting a subject search to the information provided by subject cataloguing only, neglects much other relevant information already present.

Projects linking local classifications and thesauri to the overall UDC are also under consideration. These projects are usually more difficult to realize than concordance projects of shelf numbers and UDC, especially when there are many hierarchical levels.

An experiment by the Computer Sciences department is exploring the possibilities of semi-automatic or automatic classification by UDC. This experiment investigates which fields of a bibliographic record can be used and how they should be interpreted to classify a document automatically or semi-automatically in the most appropriate UDC subject category.

Last but not least, the conversion of our DOBIS/LIBIS catalogue to Aleph 500 (MARC21 format), which will be completed in 2005, has many consequences for subject cataloguing and subject search. Our aim is to formalize the results of our experiments and the vendor's experience in our new OPAC. We have high expectations, which can be summarised as follows: classification will be considered as a valuable subject retrieval tool, subject search will be easy to use (search by words, both for classifications and pure word systems) and easily accessible ("all fields" search includes automatically a search in all subject fields), the subject information in the bibliographic record (classification descriptors and subject headings) will be clearly presented, easily understandable and will be clickable to retrieve documents with the same subjects, subject information can be shown in its semantic context (broader, parallel and narrower terms) and it will be easy to navigate through this subject structure.

It is not yet clear in detail how all our wishes will be fully realized. The Swiss NEBIS catalogue [http://www.nebis.ch] is an interesting example of the integration of a multilingual UDC search tool in an OPAC. Even when we do not necessarily want the same solution, this leaves us optimistic that we will be able to find a good solution for subject search in our catalogue.

Notes:

[1] This article is partly based on Schallier (2004).

[2] [http://www.bib.kuleuven.ac.be]. Accessed 25/08/2004.

[3] Percentages as of 4 September 2003.

[4] Percentages as of 4 September 2003. There are obvious reasons for the low percentage of MeSH: biomedical publications are a very specific part of the collection; they are often

articles; more frequently, they are available exclusively in electronic form and, as such, are not (yet) included in our library catalogue.

[5] Beyens, Werner. (2002). *E-services: een behoeftenonderzoek door middel van een webenquête bij academici aan de K.U.Leuven naar nieuwe elektronische diensten*, Leuven: uned.
[http://www.bib.kuleuven.ac.be/bibc/leeszalen/vakbib/eserviceonderzoek/eserviceonderzoek.html]. Accessed 25/08/2004.

[6] Class 2 in our authority file is based on several editions of UDC. In 2000 there was a total revision of class 2. However, we did not integrate this revision in our authority file, since the collection of our library of Theology is focused on Christian religions. UDC is also used for shelf numbers, and a revision of the authority file would therefore ask a complete reorganization of the library. For the revision of 2000, see *Extensions and Corrections to the UDC* [http://www.udcc.org/ec22_2000.htm]. Accessed 25/08/2004.

[7] I will discuss this use of UDC for building virtual library shelves in another article.

[8] [http://librisource.libis.be/LS/LS]. Accessed 25/08/2004.

[9] More specifically, Dirk Kinnaes who developed the algorithm and Tom Vanmechelen who designed the interface.

[10] UDC numbers like 563.4 Spongiaria, of which the broader term is 562 and not the expected 563, have already been signalled by Buxton (1990), Riesthuis (1998²) and Slavic (2003).

[11] [http://www.bib.kuleuven.ac.be]. Accessed 25/08/2004.

References:

Buxton, A.B. (1990), Computer searching of UDC numbers. In: *Journal of Documentation*, 46 (3), pp. 193-217.

Hunter, E.J. (2002), Classification made simple. 2nd ed. Aldershot: Ashgate.

Koch, T. (1997), The role of classification schemes in internet resource description and discovery. [http://www.ukoln.ac.uk/metadata/desire/classification/]. Accessed: 25/08/2004.

Magrijn, H.; Pontzen, S.A.Th.M.; Riesthuis, G.J.A. e.a. (ed.) (2000), Woordsystemen. Theorie en praktijk van thesauri en trefwoordsystemen. 2nd rev. ed. Den Haag: Biblion Uitgeverij.

McIlwaine, I.C. (2000), The Universal Decimal Classification: a guide to its use. The Hague: UDC Consortium.

Miksa, F. (1983), The subject in the dictionary catalog from Cutter to the present. Chicago: American Library Association.

Riesthuis, G.J.A. (1998), Zoeken met woorden: hergebruik van onderwerpsontsluiting. BBI-reeks 6. Amsterdam: Leerstoelgroep Boek-, Archief- en Informatiewetenschap Universiteit Amsterdam.

Riesthuis, G.J.A. (1998²), Decomposition of UDC-numbers and the text of the UDC Master Reference File. In: el Hadi, W.M.; Maniez, J.; Pollitt, S. (1998), Structures and relations in knowledge organization, Proceedings of the 5th International ISKO Conference, Lille, 25-29 August 1998.

Schallier, W. (2004), On the razor's edge: between local and overall needs in knowledge organization (pp. 269-274). In: McIlwaine, I.C. (ed.) (2004), Knowledge Organization and the Global Information Society, Proceedings of the 8th International ISKO Conference, London, 13-16 July 2004. London: Ergon Verlag.

Slavic, A. (2003), UDC implementation: from library shelves to a structured indexing language. 69th IFLA General Conference and Council, Berlin, 1-9 August 2003 [http://www.ifla.org/IV/ifla69/papers/032e-Slavic.pdf]. Accessed: 25/08/2004.

Taylor, A.G. (1999), The organization of information. Library and information science text series. Englewood, Colorado: Libraries Unlimited.