Implementation of a UDC-based multilingual thesaurus in a library catalogue: the case of BiblioPhil

[short paper]

Victoria Frâncu

Central University Library of Bucharest (Romania)

Cosmin-Nicolae Sabo

Universitatea de Nord Baia Mare (Romania)

The full paper is published in *Knowledge Organization*, 37, 3 (2010), pp. 209-215

1. Introduction

In recent years, information seeking behaviour has changed significantly. The development of the Internet searching services which have recently converged with interactive and user-friendly Web 2.0 applications has raised users' expectations. Library catalogues strive to improve their services to bring them up to date and closer to the functionality users are accustomed to find on the Web. The centralized library catalogue model has been replaced by a distributed cooperative model. Integrated library systems dynamically exchange information with other online systems and link to metadata repositories for additional information.

Along with this changing information environment, the change in the attitude of the information seeker is an undeniable fact. The users of library catalogues expect integrated library systems to be as flexible as Google, wikis or blogs or other sorts of online tools they are familiar with. In libraries we have a wealth of information embedded in classification notations but trying to get end-users to understand how to work with classification notations in searching for relevant information is not an option. It is generally accepted that the best approach is to make the information embedded in classification notations more easily accessible rather than to abandon the old classified catalogues and waste large amounts of valuable intellectual effort put into it. The paper describes an approach to improving classification-based subject access in a library catalogue. In order to enhance the use of Universal Decimal Classification codes in information retrieval, classification codes are represented with thesaurus terms. These terms are implemented in the bibliographic database in an automated way and used to interface classification.

2. BiblioPhil, a library system able to accommodate a UDC-based multilingual thesaurus

BiblioPhil is an integrated library system which offers the user the possibility to visualize the bibliographic records without necessarily searching the catalogue lists. Browsing rather than searching is one of the key functions of BiblioPhil.

The system provides access to bibliographic information based on a number of sorting and organising criteria such as: title (including series title), author, publisher, place of publication, year of publication, ISBN/ISSN, UDC codes and subject headings. Full text searching in the bibliographic database is enabled for the most frequently used fields: title, author, publisher, year of publication, subject headings. This prevents search failures and helps users not familiar with, e.g., subject heading lists. Truncation and Boolean operators are also available.

Complex searches are also possible by the selection of simultaneous filters according to search principles such as: the information should be a certain string of characters, or contain a string of characters, or begin with a particular string of characters. The content of the electronic catalogue is a dynamic one, almost every element of the bibliographic records being a link to all other elements which contain the same type of information.

2.1. The UNIMARC format for UDC and subject authority data

In order to support accuracy in information transfer and exchange, the structure of the authority records is based on UNIMARC. The reason for the selection of UNIMARC format is the fact that it allows easy integration in any library system. The UNIMARC UDC data field 675 is of most concern for our development. All thesaurus records are considered as separate authority records and are treated as such following the rules stated in the UNIMARC Manual (IFLA, 2008).

According to the manual, field 675 contains the UDC number or range of numbers associated with an authority heading. The UDC number may be accompanied by terms that identify the UDC number. Considering these demands and in order to achieve our objectives, field 675\$c (Explanatory terms) was redefined. The first characters in this field indicate the language of the descriptor. Since the hierarchical relations are of major importance for structured information retrieval, a special algorithm was created for automatically defining these relations. This function will further allow the expansion and restriction of the search results.

Figure 1 – Example of thesaurus search display in Winisis

MARC XML format (<http://www.loc.gov/standards/marcxml/>) was chosen as it enables data analysis and can easily be converted into other formats without data loss. In principle MARC XML framework is a simple XML schema which contains MARC data. Its main advantage is the conversion of data without loss.

3. The multilingual thesaurus benefits and limitations

The information retrieval system described in this paper is based on some of the findings of the study entitled "Multilingual access to information using an intermediate language" (Frâncu 2003). The multilingual thesaurus (Thes) used for this purpose was built according to the Guidelines in ISO 5964 (1985), the classification codes and the constituent parts of compound notations being considered as the basis for search terms in the multilingual information retrieval. Furthermore, the thesaurus was derived from and keeps the logical structure of the Pocket Edition of the UDC (BSI 1999). The verbal equivalents, descriptors and non-descriptors, are used to provide additional access points to the information in the classified catalogue and are given in Romanian, English and French.

Among the findings of this study considered relevant to the present project the following can be mentioned:

1. postcoordinated searching via multilingual descriptors derived from the UDC captions is possible
2. the use of UDC-based thesaurus terms gives better search results than the manually assigned descriptors
3. manual indexing is more exposed to inconsistencies than automatic indexing using UDC-based descriptors
4. the larger the coverage of the thesaurus the greater the occurrence of homonymous terms and hence the necessity of disambiguation
5. free-text searching produces high recall rates which might satisfy some users, while thesaurus-based searching is likely to give more precise results.

These proved to be favourable features of the hybrid information retrieval language created to serve our purposes. The accuracy of the classification codes and the terminological richness of their captions turned into thesaurus descriptors give a combination that can be successfully used in searching and retrieval. In addition, the application of this thesaurus in automated indexing show an advantage over manually assigned descriptors, as has been determined in the earlier mentioned 2003 study.

Nevertheless, integrating such a thesaurus in a bibliographic database also has a number of limitations such as: the specificity of the thesaurus has to be aligned with the specificity of the classified catalogue in order to prevent information loss; consecutive numbers connected by stroke need be managed in a special way; and polychierarchy and polysemy need to be resolved and are important issue particularly in universal disciplinary classification such as UDC in which the same or similar concept may occur in more than one class/discipline.

All these problems required additional implementation effort in order to prevent information loss.

4. The implementation of the thesaurus in the classified library catalogue

Prior research done in the field of UDC-based postcoordinate searches has been taken into account (Frâncu 1996, Riesthuis 1997, 1999). The relationship between the UDC and a word system such as this multilingual thesaurus was regarded in our case as an example of interoperability

between different types of vocabularies. The implementation process was started by defining the structures for the uploading of the authority records. Next, the selected fields were redefined and so were the algorithms for the automatic identification of broader and narrower domains. In so doing, one of the most difficult elements was represented by finding a method of searching for the bibliographic records to correspond to our criteria.

Difficulties appeared in certain instances such as mapping the series of consecutive UDC numbers symbolized by a stroke onto their corresponding descriptors. In such an instance, the individual numbers encompassed by the stroke are hidden. The algorithm developed for the present project gives as a solution an expanded series of individual numbers containing every possible number encompassed by the range given in the tables. In other words, the result is an extended authority record which contains all the individual elements encompassed by the range. Each and every hidden number will then be retrievable.

Another instance which might present some difficulties is the existence of permuted elements of a complex notation in the catalogue, e.g. (498)342.4 – Romania – Constitution, 342(498).4 – State – Romania – Constitution, 342.4(498) – Constitution – Romania (Clasificarea Zecimală Universală 1998). The first and last variants are easier to make a quick retrieval possible, but the middle one requires more elaborate processing for information retrieval. This number has to be transparently processed in two steps: 342(498) and 342.4(498). The alternative is a more complex analysis but not as elaborate as the previous case, i.e. by dividing the UDC record into main and auxiliary numbers and then considering the result expressed by the simple record 342.4(498). In practice, the syntax of the precoordinated UDC compound number, as long as it is processed and converted to descriptors, has no influence on the retrieval by words.

The instances described above are just examples of the many complex situations likely to occur in the process of defining the UDC records and their mapping to descriptors. Pre-processing of the UDC complex notations is done to allow for a simpler and more efficient information retrieval. Pre-processing means the intermediary storage of complex UDC data in order for it to be processed element by element and subsequently mapped onto thesaurus terms. In this way the information retrieval becomes a far simpler approach.

5. Information retrieval and display of search results

One of the particular features of BiblioPhil system is that the whole process of information retrieval remains “behind the scenes”. The authority files are completely responsible for the information search and retrieval, and it will not be surprising if the descriptors in the three languages are not shown in the bibliographic records. The core element that is mandatory to be present both in the bibliographic record and in the authority record is the UDC notation in two instances: one that is assigned to the bibliographic record during the indexing process, and another one in the subject authority file, i.e. in the thesaurus.

The display of the search result after querying the system may seem unfamiliar. In spite of this, its functionalities are helpful for a well structured information retrieval. The example in Figure 2 shows the descriptors associated with the query “Marriage”.

Domenii asociate cautării "marriage"			
Romana	English	French	
UDC: 173 Avort. Căsătorie (etică). Datorile copiilor față de părinți și de alți copii. Datorile părinților față de copii. Datorile soților. Etică. Morale familială. Poligamie și monogamie (etică). Viață de familie (etică). Viață de familie (etnologie).	Abortion. Duties of children to parents and siblings. Duties of parents to children. Duties within marriage. Ethics. Family ethics. Family life (ethics). Family life (ethnology). Family solidarity. Marriage (ethics). Polygamy and monogamy (ethics).	Specific domain	General domain Related domains View related books
UDC: 173.1 Căsătorie. Căsătorie (etică). Căsătorie (etnografie). Divorț (etică). Dreptul familiei. Morale familială.	Divorce (ethics). Family ethics. Family law. Indissolubility. Marriage (ethics). Marriage (ethnography). Matrimony.	Specific domain	General domain Related domains View related books
UDC: 265 Binecuvântare. Biserica creștină în general. Botez (religie). Confirmare. Căsătorie (religie). Euharistie. Hirotonisire. Penitență. Sacramente. Ultima cununecătură.	Baptism (religion). Blessings. Christian church in general. Confirmation. Eucharist. Extreme unction. Marriage (religion). Ordination. Penance. Sacraments.	Specific domain	General domain Related domains View related books
UDC: 265.5 Căsătorie (etnografie). Căsătorie (religie). Sacramente.	Marriage (ethnography). Marriage (religion). Sacraments.	Specific domain	General domain Related domains View related books
UDC: 314 Compoziția și distribuția populației. Căsătorie (demografie). Demografie. Demometrie. Familii. Fluctuații ale populației. Migrații. Mortalitate.	Composition and distribution of the population. Demography. Demometrics. Families. Marriage (demography). Migrations. Mortality. Nativity.	Specific domain	General domain Related domains View related books

Figure 2 – Example of display of thesaurus search results

The descriptors are collocated under different UDC numbers and listed in all three languages of the thesaurus. The BTs, NTs and RTs will add value to the search options and hence enrich the potential of the system through a clear representation of the semantic relationships.

The tab situated at the right corner of the display format shows the documents which correspond to the selected search term. Figure 3 shows an example of one of the retrieved documents.

Scrieri Caterinei : sfaturi unei tinere casatorite	
Autor	Shedd, Charlie W.
Editor	(Autor) Shedd, Charlie W. / (trad.) Coman, Garoafa / (trad.) Coman, Constantin, pr.
Vedeta Subiect	casatorie
Editura	Editura Bizantina
Mentiune Responsabilitate	Charlie W. Shedd ; Traducerea: Preot Constantin Coman, Garoafa Coman
An	[2006]
ISBN ISSN	973-98521-1-4
Limba	rum
CZU	265.5 173.1
Descriere Fizica	143 p.
Divizionara	2
Localitate	Bucuresti
Mediu	Carte tiparita
Colectie	Seria Casatoria - Familia

Figure 3 – Example of document retrieved as search result in BiblioPhil (query: marriage)

Another example of a search query used for demonstration purposes is expressed by a phrase, i.e. "international painting robbery". For this query the UDC notation is 343.71:75(100) and its component elements are (100) as location, and 343.71 and 75 as main numbers. The application of the specific algorithms resulted in a set of retrieved documents of which one is displayed in Figure 4.

Operatiunea linz : cel mai mare jaf de opere de arta al tuturor timpurilor	
Autor	Seydewitz, Ruth Seydewitz, Max
Editor	(Autor) Seydewitz, Ruth / (Autor) Seydewitz, Max / (trad.) Raus, Ileana / (pref.) Dragut, Vasile
Vedeta Subiect	arte plastice / grafica / infractiuni / infractiuni impotriva proprietatii / istoria artei
Editura	Meridiane
Mentiune Responsabilitate	Ruth Seydewitz, Max Seydewitz ; traducere de Ileana Raus ; prefata de Vasile Dragut
An	1979
Limba	rum
CZU	73/76(100)"1939/1945" 343.711:73/76(100)"1939/1945" 94(100)"1939/1945" 93(100)"1939/1945"
Descriere Fizica	276 p. : il.
Divizionara	7177
Localitate	Bucuresti
Mediu	Carte tiparita
Serie	259
Colectie	Colectia Biblioteca de arta. Biografii, memorii, eseuri

Figure 4 – Example of document retrieved search result in BiblioPhil
(query: international painting robbery)

What is remarkable here is the difference in specificity between the UDC notation given by the indexer and the UDC elements in the authority record that are used as a basis in the search process. In spite of this difference the document was retrieved as a result of the algorithms applied in order to enable the search.

6. Conclusions and future developments

The solution proposed by this project implies a significant change in the OPAC display, breaking with tradition and offering a completely new and more attractive visual arrangement (<http://www.bibliophil.ro/UDCresearch>). More similar to the way that Internet search results are shown, this arrangement shows very clearly and efficiently which are the bibliographic resources that best fit the search criteria used to interrogate the system. The simple fact that the classification notations formerly used in indexing are no longer necessary to find the indexed information is beneficial for both the indexer and the end user. By means of the reciprocal relations between the UDC codes and the derived descriptors, and between those and the nondescriptors, the users of the system are permanently aware of the form of the access points available for information retrieval. The solution saves the time of the indexer and provides more user-friendly and easier access to the bibliographic information. At the same time, the multilingual aspect of the UDC-based thesaurus enhances information access by bringing in a greater number of online users.

In addition to the above characteristics of the system, a list of general subjects has been built and is available in Romanian, English and French. This list is capable of functioning as a search tree structure for quick searches and thus it can provide a rapid overview of the main disciplines existing in the library collection. Hence, the disciplines can be further subdivided into more specific domains, enabling a deeper exploration of the bibliographic resources held by the library.

References

- BSI (1999) *Universal Decimal Classification Pocket Edition*, PD 1000. London, British Standard Institution
- Clasificarea Zecimală Universală (1998) *Clasificarea Zecimală Universală: ediția medie internațională în limba română, Partea I: Tabele sistematice*. București: Biblioteca Națională a României.
- Frâncu, V. (1996) Building a multilingual thesaurus based on the UDC. In: Green, R. (ed.) *Knowledge Organization and Change: Proceedings of the fourth conference of the International Society for Knowledge Organization (ISKO)*, Washington DC, 15-18 July, 1996. Frankfurt/Main: Indeks Verlag, pp. 144-154.
- 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*. PhD thesis, Faculteit Taal- en Letterkunde, Germaanse Taal- en Letterkunde, Universiteit Antwerpen. Available at: <http://dlist.sir.arizona.edu/1862/> [accessed 25.07.2009]
- International Organisation for Standardization (1985) ISO 5964. *Guidelines for the establishment and development of multilingual thesauri*. Geneva: ISO.
- Riesthuis, G. J. A. (1997) Decomposition of complex UDC notations. In: McIlwaine, I. C. (ed.) *Knowledge Organisation for Information Retrieval: Proceedings of the Sixth International Study Conference on Classification Research*, University College London, 16-18 June 1997. The Hague, FID, pp. 139-143.
- Riesthuis, G. J. A. (1999). Searching with words: re-use of subject indexing. *Extensions and Corrections to the UDC*, 21, pp. 24-32.