Piros Attila
Nebu Hungary Kft.
atilla.piros@gmail.com

A NEW APPROACH TO UNIVERSAL DECIMAL CLASSIFICATION AS THE INDEXING LANGUAGE FOR A MECHANIZED INFORMATION RETRIEVAL SYSTEM

Preface

Universal Decimal Classification (UDC) is one of the world's foremost library classification systems, being used in hundreds of thousands of library collections and their online catalogues, which try to adapt to the particularities of the system in order to retrieve contents based on their subjects.

Despite its prevalence the information retrieval systems being currently used are unable to thoroughly utilize all the facilities following from its analytico-synthetic and faceted nature.

In this paper I will try to outline a possible way to describe the inner structure of UDC compounds in a markup language. I will also demonstrate a software interpreter which is able to convert UDC numbers into the format mentioned immediately above. This conversion and its result might be used as a basis of more effective methods to retrieve information employing UDC.

About the Usage of Universal Decimal Classification in Information Retrieval Systems

About Universal Decimal Classification

UDC is one of the largest and most widely used classification systems in the world. Since the publication of its first edition more than one hundred years ago its different editions have been published in no fewer than 39 languages and it is used in at least 124 countries (Slavic 2008); there are currently over 140,000 collections which have been indexed using UDC in Europe alone (Slavic 2012).

Its widespread use is partially due to the radical innovations made by its inventors, creating the first analitico-synthetic classification („UDC History").

The other key to its prevalence is its having been under continuous review by a non-profit consortium in order to ensure that it is constantly kept up-to-date and reflects the current status of the sciences and the changes in the ways it is used. The UDC Consortium publishes the officially authorized changes to UDC annually in its official journal („Extensions & Corrections to the UDC").

Automation of Using UDC

UDC was most likely used as the indexing language of a mechanized information retrieval system for the first time by E. G. Brisch, who presented his experience of adaptation of UDC in a system using punched card in 1948 (Bhattacharya 1969).

Software has been approached classification from a bibliographic point of view from the very beginning; it has been trying to retrieve documents with the highest possible recall and precision based on the equality of the complete UDC notations or some parts of them. In order to reach this goal they usually handle UDC numbers as simple keywords or, sometimes, build a KWIC or KWOC type index of the parts of the numbers (cf. Rigby 1967 and Buxton 1992).

In a pilot study conducted in 2004-2005, Aida Slavic examined the existence of the following functionalities in thirty Web OPACs (the numbers following the names of the functionalities are the percentages obtained as the result of the research) (Slavic 2006):

1. automatic right truncation 66.7%
2. approximate matching (i.e. 'no zero result' option) 46.7%
3. availability of a Boolean search 10%
4. searching parts of a complex UDC number 23%
5. searching a UDC caption 36.7%
6. searching/browsing from an authority record choosing any related terms within the record. 16.7%

Although searching the parts of a number is also possible if numbers are handled just as simple text, flexibilities in citation order, the possibility of intercalation etc. (cf. „UDC Summary Linked Data") make it difficult. To avoid losing data the searcher should collect every possible citation order of a compound. Using wildcards is a useful tool in doing this as it raises the level of recall, although this causes the level of precision to be lowered. Handling consecutive extensions and special auxiliaries effectively is impossible in this way. Permuting the numbers and building a KWIC-index may raise the level of recall but doesn't solve the problems mentioned above.

KWOC (Keywords out of Context) form indexes were used for UDC numbers by Klaus Schneider and Karl-Heinz Koch in the Sixties. They compiled an index of parts of complex concepts by permuting them and isolating from the other index parts in order to retrieve documents with greater precision (Rigby 1974, 36). This method primarily supports post-coordinated searches, which means that the user is allowed to create complex queries from simple UDC numbers using Boolean operands. Just as in the case of using numbers as keywords, retrieving numbers hidden in consecutive extensions or searching for special auxiliaries is not supported in this way, nor is handling subgrouping. The differentiation of concepts like 329.17:329.12 (relations of national and liberal movements) and 329.17'12 (national-liberal movements) is also impossible in this way.

The problem of the methods mentioned above may derive from the fact that information regarding context is lost permanently during the indexing phase; so the system cannot utilize the information about which parts of a complex number join to each other, with which operands and in which order during the searching phase.

In the last few years managing authority files have become the predominant model of utilizing classification systems in modern information retrieval software. The main advantage of this method is that an authority record can contain not only the

classification number, but metadata regarding this, such as its broader, narrower and related classes, its equivalents in other knowledge organization systems, like thesauri, subject heading lists or other classifications, and its access points using either the codes of the classification or in natural languages. It is possible to allow the authority files to be updated automatically, which helps to keep the catalog up-to-date. In addition to the capacity to use the same authority records many times, the main advantage of utilizing authority files is that the indexers and searchers are able to access subjects by using their descriptions in a national language instead of the artificial codes of the classification systems. In addition to the standard MARC authority formats special authority formats have also been created for classification authority files, like MARC 21 Format for Classification Data („MARC 21 Format for Classification Data") or UNIMARC Classification Format („Concise UNIMARC Classification Format") for example (Slavic 2007).

Authority control brings the possibility of a restricted, pre-coordinated way of retrieving information based on approximate matching into existence and also renders it possible to utilize the pre-defined relationships between the concepts – such as thesauri and any other controlled vocabularies.

In spite of its many advantages, employing UDC authority files restricts deeper content analysis and the usage of synthesized UDC numbers, because a new authority record must be compiled for every newly created concept; if it is not possible, the indexer has to use one of the existing concepts. Access to authority records is also restricted because the access points and relationships are predefined. If the searcher prefers to compile his/her own UDC compound which meets his/her own requirements instead of browsing or selecting a keyword, he/she will access the authority records through the same methods of using keywords and KWOC indexes as in an average OPAC, which results in the disadvantages mentioned above.

A Feasible Alternative

In addition to the above, the system of the classification, thanks to its analitico-synthetic nature, makes another approach feasible. This means that UDC could be employed as the indexing language of systems using more complex retrieving algorithms, which would utilize linguistic analysis of UDC numbers instead of the currently used methods based on the exact matching of notations. However, in order to reach this goal we need more sophisticated methods to interpret UDC notations, to discover their inner structure, containing its parts, their roles and the way they join to each other.

The markup languages provide us with a useful tool to describe the inner structure of UDC numbers with necessary elaboration. After parsing the numbers and interpreting them in a markup language, the results of this process can be used to build a database of UDC numbers, which can be utilized as the base of more complex search engines based on the vector space model for example.

The first step in reaching this goal is to create a schema definition which will provide us with the confines of the description of numbers and a software interpreter which will automatically translate the numbers into the markup language based on the schema. The next steps will involve solving storing the results of the parsing process onto a database,

and, finally, devising and implementing the searching algorithms which make it possible to retrieve information from the database.

Aside from the second phase, the results of the first phase may satisfy the needs of other applications after further transformations. For example, the KWOC-form index entries could be created automatically just as descriptions of the numbers in MARC-formats, which could be utilized by OPACs and authority files being currently used.

The subject of this lecture is to present an XSD schema and a software interpreter utilizing this, which makes it possible to interpret any UDC-number automatically into an XML format, without losing any relevant information about its parts and structure. The interpreter analyzes UDC numbers in a syntactic way using only the rules of UDC, which means that the program doesn't need to store UDC numbers or any part of the tables.

The software is under development; its current version is available online for trying and testing purposes in the URL: http://interpreter-eto.rhcloud.com.

Analysis of UDC Numbers[1]

The Structure of UDC Compounds

In this chapter I will demonstrate the way a UDC compound is built using the following example:

[323.272+323.83]"1848/1849"(439):821.511.141-1"18"

Description (English): Relationship between the Hungarian revolution and war of independence of 1848/1849 and the Hungarian poetry in the 19th Century.

Description (Hungarian): Az 1848-49-es forradalom és szabadságharc és a tizenkilencedik századi magyar irodalom kapcsolata.

The hierarchical structure of this number can be listed by taking account of the rules of joining the parts to each other and the order of precedence of the auxiliary signs:

[323.272+323.83]"1848/1849"(439):821.511.141-1"18"
└relation, notation: [323.272+323.83]"1848/1849"(439):821.511.141-1"18"
└subgrouping, notation: [323.272+323.83]"1848/1849"(439)
└addition, notation: 323.272+323.83
└main table number, number: 323.272
└main table number, number: 323.83
└common auxiliary of time, number: „1848/1849"
└common auxiliary of place, number: (439)
└ main table number, notation: 821.511.141-1"18"
number: 821.511.141
└special auxiliary subdivision, number: –1
└common auxiliary of time, number: „18"

[1] The rules and principles outlined in this chapter and used during the implementation of the presented software came from the related parts of the Hungarian UDC Edition published in 2005 („Egyetemes Tizedes Osztályozás") and from the scope notes published on UDC Summary page („UDC Summary"). The example was also built by using the editions mentioned immediately above.

We can see the relevance of relations between the parts of the number, which means that recognition of the parts themselves is not satisfactory without saving the way they join to each other.

The Hardness of Processing UDC Numbers

Whilst interpreting UDC numbers we can meet problems caused by the inconsistency of the rules of the classification itself.

For instance, Pauline Atherton and Robert Freeman, the supervisors of the first project which investigated the possibility of using UDC in mechanized systems, „noted a number of problems with handling UDC notation:

1. Some of the notational devices serve only for visual convenience (i.e., the decimal point).

2. Some of the notational devices use the same punctuation symbol. The order of sorting these causes problems when using a computer.

3. Recognition of some devices requires examining two characters, for example distinguishing = and (=.

4. .0 and .00 incorporate meaningful use of the decimal point, which elsewhere is only for convenience." (Buxton 1992)

The preceding versions of UDC contained special auxiliaries without indicator characters, being almost absolutely unidentifiable. Direct alphabetical specifications don't have indicator characters too, and the set of characters can be contained by them is undetermined (unless we restrict the rules of transcription, which would make the interpreter being confined to a few languages).

Continuous revision is one of the biggest advantages of UDC. Nevertheless, it makes implementing interpreter and information retrieval software especially hard, because they should be capable of interpreting the earlier and the present rules, as well as integrating future modifications (cf. „Major Changes to the UDC 1993-2013").

The Xml Schema Definition (XSD)

The output of the interpreting process is the presentment of the UDC number in Extended Markup Language (XML), which follows the hierarchical structure described above. The basis of the presentment is an XML Schema Definition (XSD), which is constructed to try to make it possible to describe any regular UDC number.

In addition to building XML representations, the XSD can be utilized for the validation of UDC numbers, which means that any XML parser program is able to recognize if the interpretation was built from an invalid UDC number (if the interpreting process itself didn't recognize the problem whilst it was parsing the number).

The complex types of the XSD describe the possible elements of UDC numbers: the numbers of the schedules, the auxiliary signs and the UDC number, while the simple types of the XSD were introduced for validation purposes.

In accordance with the schema an XML must contain one and only one element to describe the UDC concept; this element contains the other elements as we can see in the hierarchy above. It also contains the notation of the number and the UDC edition which was used to built it as attributes and the descriptions of the concept in different languages as elements too.

The complex type cited above describes every possible form of the UDC numbers which observes the rules of UDC.

An Example of the Presentment of UDC Numbers

The result of processing the UDC compound analyzed in the chapter about *The Structure of UDC Compounds* can be seen below:

<?xml version=„1.0” encoding=„UTF-8”?>
<udc:udc_concept xmlns:udc=„http://www. inf.unideb.hu/library/udc/xml” xmlns:xsi=„http://www.w3.org/2001/XMLSchema-instance” udc_edition=„2014” notation=„[323.272+323.83] „1848/1849”(439):821.511.141-1”18”„>
<udc:description xml:lang=„EN”>Relationship between the Hungarian revolution and war of independence of 1848/1849 and the Hungarian poetry in the 19th Century</udc:description>
<udc:description xml:lang=„HU”>Az 1848-49-es forradalom és szabadságharc és a tizenkilencedik századi magyar irodalom kapcsolata</udc:description>
<udc:main_table_relation>
<udc:main_table_subgrouping>
<udc:main_table_addition>
<udc:main_table_number number1=„323.272”/>
<udc:main_table_number number1=„323.83”/>
</udc:main_table_addition>
<udc:common_auxiliary_independent xsi:type=„udc:common_auxiliary_of_time”>
<udc:common_auxiliary_of_time_number number1=„„1848”„ number2=„„1849”„/>
</udc:common_auxiliary_independent>
<udc:common_auxiliary_independent xsi:type= „udc:common_auxiliary_of_place”>
<udc:common_auxiliary_of_place_number number1=„(439)”/>
</udc:common_auxiliary_independent>
</udc:main_table_subgrouping>
<udc:main_table_number number1=„821.511.141”>
<udc:special_auxiliary xsi:type=„udc:special_auxiliary_number_hyphen” number1=„-1”/>
<udc:common_auxiliary_independent xsi:type=„udc:common_auxiliary_of_time”>
<udc:common_auxiliary_of_time_number number1=„„18”„/>
</udc:common_auxiliary_independent>
</udc:main_table_number>
</udc:main_table_relation>
</udc:udc_concept>

The XML contains a UDC concept containing the notation and the UDC edition used to built it as attributes as well as its subconcepts as elements. The XML also contains the descriptions of the concept in different languages.

The concept contains an auxiliary sign (or operand) which is a relation of an algebraic subgrouping and a main table number; the subgrouping contains an addition of two main table numbers and two common auxiliary numbers (one of place and one of time) joining it; the second operand of the relation is a main table number specified by a

special auxiliary number and a common auxiliary of place. Each of the elements listed above appears in the proper place in the hierarchy.

Conclusions

As Ágnes Hajdu Barat points out in the introduction to her study about the possibilities of Hungarian OPACs „there is marked interest in the UDC's potential to assist growing numbers of Internet users. The UDC can play a role of integration in knowledge organization." Her conclusion is that we should keep UDC in our retrieval systems instead of abandoning it (Hajdu Barat 2006).

The fact, that questions such as the above have even been raised reveals that the information retrieval software systems currently used haven't reflected the significance of UDC numbers by using all the capabilities of the classification. However, further research and improvements can result in an increase in the effectiveness of bibliographic analysis and retrieving information by taking advantage of the analitico-synthetic and faceted characteristics of UDC.

The markup languages, and software frameworks for using them, may provide a useful tool to support the improvements mentioned immediately above. That is why I decided to publish the principles above and to implement the prototype of the presented software interpreter.

Bibliography

Bhattacharya, G, *Vital role of depth classification in a system for document-finding: a trend report*, Library Science with a slant to Documentation, Vol. 6 (1969), iss. 1, 52-70.

Buxton, A., *Computer Searching of UDC Numbers*, Encyclopedia of Library and Information Science, 51 (1992)

Hajdu Barát, Á., *Usability and Responsibility*, Extensions & Corrections to UDC, Vol. 28 (2006), 46-55.

International Federation of Library Associations and Institutions, *Concise UNIMARC Classification Format (20001031)*, http://archive.ifla.org/VI/3/p1996-1/concise.htm (accessed on August 23, 2014)

Library of Congress Network Development and MARC Standard Office, *MARC 21 Format for Classification Data*, http://www.loc.gov/marc/classification/eccdhome.html (accessed on August 23, 2014)

Egyetemes Tizedes Osztályozás [Universal Decimal Classification] : UDC Publ. No. P057. Budapest: Országos Széchényi Könyvtár Könyvtári Intézet, 2005.

Rigby, M., *Computers and the UDC; A Decade of Progress 1963-1973*. The Hague: International Federation for Documentation, 1974.

Slavic, A., *The level of exploitation of Universal Decimal Classification in library OPACs: a pilot study*, Vjesnik bibliotekara Hrvatske, Vol. 49 (2006), iss. 3-4, 155-182. http://arizona.openrepository.com/arizona/handle/10150/105346 (accessed on August 23, 2014)

Slavic, A., Cordeiro, M. I., Riesthuis, G., Enhancement of UDC data for use and sharing in a networked environment, Paper based on the talk presented at The 31st Annual Conference of the German Classification Society on Data Analysis, Machine Learning, and Applications, March 7-9, 2007, Freiburg i. Br., Germany, http://arizona.openrepository.com/arizona/handle/10150/106330 (accessed on August 23, 2014)

Slavic, A., *Use of the Universal Decimal Classification: A World-Wide Survey*. Journal of Documentation 64(2) (2008): 211-228.
Slavic, A., *UDC libraries in the world – 2012 study*. Universal Decimal Classification Blog (blog), August 20, 2012. http://universaldecimalclassification.blogspot.hu/2012/08/udc-libraries-in-world-2012-study.html (accessed on August 23, 2014).
UDC Consortium, *Extensions & Corrections to the UDC*. http://www.udcc.org/index.php/site/page?view=ec (accessed August 23, 2014).
UDC Consortium, *Major changes to the UDC 1993-2013*. http://udcc.org/index.php/site/page?view=major_revisions (accessed August 23, 2014).
UDC Consortium, *UDC History*, http://udcc.org/index.php/site/page?view=about_history (accessed on August 23, 2014)
UDC Consortium, *Universal Decimal Classification : Summary*, http://www.udcc.org/udcsummary/php/index.php (accessed on August 23, 2014)
UDC Consortium, *UDC Summary Linked Data: Common auxiliaries of place. Table 1e,* http://udcdata.info/001951 (accessed on August 23, 2014)