Decomposition of UDC-numbers and the text of the UDC Master Reference File

Abstract: There exist in the world many bibliographical databases, which are indexed with the UDC. Searching in the databases is difficult for everyone not very experienced in the use of this classification scheme. The notations can be very complex and often it would be desirable to be able to search on only a part of the notation. In this paper algorithms for the decomposition of complex UDC-notations and adding a description to the resulting simple notations are discussed. In the second part of the paper the resulting texts are discussed.

1. Introduction

There exist in the world many bibliographical databases, which are indexed with the UDC. UDC-numbers as given to the documents, which are described in these databases, consist of one or more numbers taken from the main table, often combined with general and special auxiliaries. The resulting subject notations can be quite complicated. For many users of databases indexed with UDC searching is not very easy and often the results of the search actions are not what is expected, because the user failed to use the correct subject notations. How can an user find out that he should use 649.1:616.28-008.13/14-053.2 for documents on domestic care for deaf and hard-hearing children? This complicated UDC-number consists of more parts: the main numbers 649.1 *Domestic childcare* and 616.28 *Ears. Otology*, the special auxiliaries -008.13 *Distorted hearing* and -008.14 *Deafness* that belong to 616.28, and the general auxiliary -053.2 *Children*.

Searching would be easier if the user could search with the separate parts, and even more easier if the user does not to have use UDC-notations at all. This goal can be reached if we are able to break the complex UDC-numbers in the composing parts and to replace each part with a text, giving the meaning of the numbers.

In this paper a research project to achieve the goal is reported.

2. Decomposition and finding text

The UDC consists of a main table and eight general auxiliary tables, if we count alphabetical additions as a general auxiliary table. Each of these tables has its own symbol(s) that distinguish them from each other. A notation from the main table always starts with a digit (but not with a 4). The notations from the general auxiliary tables all start with a character that is not a digit. Some tables also have a special character that marks the end of the auxiliary. We can see that 316(498) consists of two parts: a number from the main table – 316, beginning with a digit – and a number from the general auxiliary table of place – (498), beginning with a ')' followed by a digit greater than 0, and ending by a ')'. In 316=135.1 the end of the general auxiliary =135.1 is found by reaching the end of the notation. The end of the main number 316 is found because the = denotes that a general auxiliary of language is starting. Notations from the tables can be combined to quite complex subject notations such as 741.5(498)Taru, E.(084), 641/642(493Bruxelles)(084) or 323.12(=411.16:498).
There are also special auxiliaries valid for a part of the main table, or a part of one of the general auxiliary tables. These special auxiliaries depend for their meaning on the main number or general auxiliary they belong to. There exists three types of special auxiliaries: beginning with .0..., with an apostrophe, or with a hyphen followed by a digit greater than 0. In a few classes there exists what is called final digits. This means that after all subdivisions of these classes digits can be added which function in the same way as special auxiliaries. Example: 539.12...13 Decay. 539.128.4 Alpha particles. 539.128.413 Decay of Alpha particles.

The meaning of a simple notation – i.e. a notation not combined with any other notation from the main table or any general auxiliary table – can in most cases be found in the tables of the MRF. Examples of such simple notation are (498) Romania, 316 Sociology, and 662.754 Organic chemicals as liquid fuels. Alcohol, methanol as fuels, but also notations with special auxiliaries such as 615.357.015 Hormones – Pharmacodynamics. General principles of pharmacology are considered to be simple. Decomposition means that subject notations – given to bibliographical records – that are not simple in this sense are broken in parts that are simple.

There is a complication: there are simple notations that can not be found in the MRF tables. There are two reasons for this phenomenon: special auxiliaries and parallel subdivision. Special auxiliaries are given at one place in the MRF and can be used for all notations lower in the hierarchical chain. Example: 314.018 give the special auxiliary .018 the meaning Demometrics (measurement techniques). This auxiliary can now be used throughout 314 Demography. 314.22 is Births. Measurement of birth rates etc. is 314.22.018.

Parallel subdivision means that the subdivision of a given notation in a table is also used for another notation. 376.1 Organization of special schools is subdivided like 371 Organization of educational and training system. School organization. To denote Optional courses at special schools 376.121.419 can be used. This notation is built of 376.1 Organization of special schools and 371.214.19 Optional courses. The part after 376.1 and the use of the point is adjusted: a point comes after each three digits.

Decomposition of complex UDC numbers has to be done in steps:
1. Subject notations containing two or more main numbers linked with a connector (;, :: or +), are split into as many notations as there are main numbers. 649.1:616.2808.13.14-053.2 is divided into 649.1 and 616.28-008.13/.14-053.2. But 32(47+57)1918/1964/0(82) and 323.12(=411.16:498) have one main number only. It is thus not enough to look for a connector between notations and then conclude that the notation has more than one main number. Connectors inside general auxiliaries do not count.

2. General auxiliaries, including alphabetical additions, are separated from main numbers. 821.135.1.09=18/19» is split into 821.135.1.09 and «18/19». The algorithm is rather simple. Only alphabetical additions sometimes require special attention, e.g. when there are brackets inside the addition as in 821.134.2Claréns(Alas Leopoldo) or digits like in 681.3Windows 95.

3. Combinations of two or more general auxiliaries from the same table are split. (47+57) is split into (47) and (57). In the notation 323.12(=411.16:498) the two general auxiliaries come from different tables: (498) is from the table for place and (=411.16) from the table for ethnic groups. The algorithm has a special provision for these cases.

These first three steps can be done without looking up anything in the UDC tables. The following steps require that an UDC-edition be consulted. For my research I used the so-called Master Reference File (MRF). Since 1993 the MRF is the standard version of the UDC. It is
available as a computer readable file and updated each year.

1. Ranges of UDC-numbers denoted with a / are broken down into the separate numbers of the range. Each potential number within the range is looked up to see if the number is an existing, valid UDC-number. 556.3/.5 gives the numbers 556.3, 556.4 and 556.5. The first and last notations of a range are always given, even when they do not exist in the MRF. Ranges can be complex: 72/73.033.4/.5 has two '/'s. It should split up into 72.033.4, 72.033.5, 73.033.4 and 73.033.5. The algorithm can not handle these cases correctly.

2. UDC-numbers with final digits are separated in the final digits and the rest. There is special algorithm for each class where final digits exist. The subject notation 7.033.4(410.1)1 is first split into (410.1) and the remaining 7.033.41. The algorithm for main numbers, followed by a special auxiliary can not find text for 7.033.41. Because this notation starts with a 7, the end digit algorithm looks if there is a special auxiliary .033…, 034… or .035… If this is the case … is added at the appropriate place. For 033… is that after the 4th digit. This gives 7.033.4…1. This notation consists in the MRF.

3. Notations, which are the result of parallel subdivision, require special treatment. There are several possibilities if a notation is not found in the MRF. One of them is that the notation is the result of parallel subdivision. The notation 616.819 is not found in the MRF. There is no way to see at the notation itself that it is the result of a parallel subdivision. The algorithm used truncates the last digit and looks for the resulting notation in the MRF. 616.81 is found and in the record for it there exists a note that 616.81 should be subdivided like 616.831. We can now put the digit 9 that we have taken off our start notation 616.819 behind 616.831 (adjusting the points) and search for 616.831.9. This gives [Diseases of the] Cerebral meninges. To find the meaning of 616.819 we combine the meanings of 616.81 and 616.831.9. This gives Central nervous system. Cerebrum, encephalon. Brain. Cerebral meninges.

4. When we try to find the meaning of a notation with a special auxiliary the problem is mostly to find where in the tables we can find the meaning of the special auxiliary. In a few cases the notation as in the record is found in the MRF. If not, the standard way is relatively simple: split the main number and the special auxiliary, take the last digit of the main number, add the special auxiliary and search again. Repeat this till the notation is found. This gives the meaning of the special auxiliary, not of the notation as a whole! To get the meaning of the total notation search with the main number alone. The two texts combined give the meaning of the notation. An example: 821.135.1.09-1 has two special auxiliaries: .09 and −1. First two notations 821.135.1-1 and 821.135.1.09 are formed. Truncation gives at least 82-1 Poetry and 82.09 Literary criticism. The main number 821.135.1 is the result of a parallel subdivision. 821 is subdivided parallel with the general auxiliary table for languages. 821.135.1 means Romanian literature.

A special algorithm is developed that searches the MRF for the text. This algorithm 'knows' the structure of the MRF and adds to each notation found the description in English as found in the MRF.

The algorithms have been used to find texts for a part of the catalogue of the Biblioteca Centrală Universitară of Bucharest (Romania) built in the period 1992-1996. This part (called BCUB) has some 71,000 records of which 52,000 contain at least one UDC notation. In total there were 104,000 notations. 7000 notations only got a text after truncation of at least one
digit. 800 received no text – for the main text or for a special auxiliary – even after truncation. The main reasons for truncations are:

- The notation does not exist at all. More than 3700 records of the BCUB have the nonexistent notation 043 used for Dissertations. As 04 also do not exist the notation truncates to 0 and the text belonging to 0 is given to the notation 043.
- The notation is lower in a hierarchical chain than the MRF allows. This is a residue from the time the Biblioteca Centrală Universitară was using a more detailed edition of the UDC than the MRF (before 1992).
- The notation is not formed according the rules of the UDC. This can mean that a character that denotes the end of a general auxiliary is forgotten, or that there is a space within a notation and many errors more.
- The notation is well formed but the result of a parallel subdivision that is not mentioned in a regular way. Normally the instruction for a parallel subdivision is given in a subfield 011 of the relevant record of the MRF, but some times the instruction is to be found in a note. The algorithm can not find such instruction.
- The notation is the result of a «double parallel subdivision». This means that the notation is subdivided parallel to a notation that itself is the result of a parallel subdivision. The algorithm can not cope with this.
- The notation is a special auxiliary that is not mentioned at the relevant main number or at a notation higher up in the hierarchical chains but elsewhere, e.g. in a note. The algorithm used does not truncate special auxiliaries. This means that in these cases the special auxiliaries get no text at all.

On the whole the result of the decomposition is good. When the subject notations given to documents are formed according to the rules, decomposition is possible in more than 99% of all cases. Only ranges with / can not always be decomposed. In many cases the use of a range could have been avoided by using a notation higher up in the hierarchical chain. It makes no sense to use the notation 025.1/9 as is done in a record of the bibliographical database. Why not used 025? The use of ranges should be restricted to those ranges that are given in the MRF.

3. The text of the UDC

Most librarians and other information specialists will see the UDC as a controlled language. This is correct when the notations as given in the tables are considered to form the vocabulary of the UDC. The notations are controlled: each notation has its own meaning. But here lies also the problem: what is the meaning of a notation. 316 by itself does not have any meaning at all. A notation is given a meaning in two ways: by the description and by the context.

Descriptions

The algorithms used can add the text of the descriptions to the notations found in the tables. The notation 316 will get without any difficulty the text Sociology.

When a notation is the result of parallel subdivision, this notation will not be found in the MRF. The algorithm for parallel subdivision however will find out that the notation 376.121.419 is the result of parallel subdivision of 376.1 like 371. The part after 376.1 will be put after 371 to form 371.214.19. This notation will give the text Optional courses that is not enough to give the correct meaning of 376.121.419. The text belonging to 376.1 Organization of special schools is added to give Organization of special schools. Optional courses.

For notations with special auxiliaries applies the same. The notation 81`23 gets as text
Linguistics and languages (for 81) and Psycholinguistics for the auxiliary ’23. When in a notation with a special auxiliary the main number is the result of a parallel subdivision then the text consists of three parts. In the notation 821.135.1-1 the main number 821.135.1 is the result of parallel subdivision of 821 Literature with =$135.1 Romanian. The special auxiliary—1 means: Poetry. Poems. Verses. It is possible that both the main number and the special auxiliary came about by parallel subdivision.

Context

For many subject notations however it is not enough to add the description to get the meaning of the notation. The auxiliary of place (498) is used for Romania and the text in the UDC is Romania. Republic of Romania. This is clear in itself. There is no reason to add the text of (49) or (4). The subdivision (498.21) has as text: Dolj (Doljiiu). Craiova. It would be preferable to add the text of (498) also to the notation (498.21). In general it is a good idea to add to notations of the general auxiliary table that denote regions smaller than a country, the text belonging to the notation for the relevant country. For the other general auxiliaries it is not necessary to add the text of any notation higher in the hierarchical chain. The text for the general auxiliary of persons -058.566 is Civil detainees. Internees. Political prisoners. There is no reason to add the text for any notation higher in the hierarchy.

To understand the meaning of a notation of the main table it is almost always necessary to add extra text. The notation 686.1 has Bookbinding as description. Here the context is necessary to know that with this notation bookbinding in the printing industry is meant. Bookbinding in a library context is 025.7. The description of 025.7 is Binding, upkeep, and repair of books. In the UDC 46 notations have the word absorption in their text, in very different context. The word nations is found in 11 places. The context can be sociology, political science or law. Natural is mentioned in the text of 168 notations, including 788.41 For natural horn. For hunting horn ….

Synonyms

Synonymy plays a role in the UDC in two ways. The meaning of a notation can often be expressed in the description with different words and phrases, but it is also possible that two notations in the table are synonyms.

In many places in the UDC these synonyms are enumerated in the descriptions. The description for 65.011.72 is Errors. Defects. Faults. Deficiencies. McIlwaine(1993) says in the official guide to the UDC: «The class description term is the essence of the class entry. It defines the concept within its hierarchical context by exactly describing it in natural language. The first, and often the only part of the description is the primary term which may be a word or phrase. This may be followed by additional terms (synonyms) expressing the concept. These additional terms are particular important when a verbal index to the classified sequence is being constructed since they suggest synonyms and alternative terms that may be sought by users. Full use should be made of them in information retrieval.»

There are also descriptions that enumerate pseudosynonyms. Sometimes a number of
related terms are given in one description, like in 626.025 Diving suits. Breathing equipment. Ancillary equipment. Also 'upward posting' is used. The description for 626.113 is Canals with locks. Summit canals. Junction canals. 347.786 has as description Dramatic arts. Plays. Theatre. Films. Screenplays, scenarios. The context is given under 347.78 Artistic and literary property. Author's rights. Copyrights. In the English Medium Edition (1993) the upward posting is denoted by the term Including: before the hierarchical lower levels, in the examples after locks and arts.

In the vocabulary – the notations – of the UDC itself synonyms can also be found. 658.23 and 725.2.011.2 both denote Layout of (commercial and business) buildings. The context – the point of view – is in theory different: 658 means Business management, administration. Commercial organization and 72 is Architecture. Yet it is not improbable that of two indexers one will use the 72-notation and the other the 658-notation for the same book or that the same indexer will use one notation today and the other tomorrow for documents which deal with the same subject.

Searching for terms in the descriptions of the UDC can give unexpected results, due to this problem of synonymy. 'Open' prisons leads to 725.63 under 72 Architecture. At 343.81 (under 343 Criminal law) the term is nowhere to find. The concept is to find under 343.84 Stages in the punitive process. Degrees of severity.

The problems with synonyms in the descriptions can be solved by thesaurification of the UDC with USE and USED FOR relations (see Riethuis & Bliedung, 1991, Francu, 1996). This is a long way. In the meantime it would be a good suggestion to add (pseudo) synonyms to the descriptions. More faceting and a more strict logic in building classes can solve the second synonymy problem. There should one place where notations for the layout of buildings are given. In other places where the layout of a building is involved a colon relation can be used.

Homonymy and polysemy

When we use the natural language of the descriptions we are searching with words. In any information system where words are the entries the question if homonymy is a problem depends on the domain. The larger the domain the greater the chance that we will have homonymy and polysemy, that the same terms are used with different meanings or in different context. For the UDC the domain is universal. Polysemy is therefore quite common in the descriptions of the UDC. The meaning of the term sandwich in Advertising on itinerant carriers. Sandwich-man (659.134.2) is far from the original meaning in 641.84 Bread with dressing or filling. Sandwiches. Bread and butter, spread etc. Seldom however the meaning of the homonymic terms is unclear, in most cases we have enough context in the description for knowing what is meant exactly.

Hierarchical relations

Hierarchical relations are the alpha and omega of classifications. Each classification subdivides classes in genera or parts. In the UDC subdivision along a hierarchical chain is mostly reflected in the notation. One hierarchical level lower means one digit more in the notation. There are exceptions. Sometimes a hierarchical lower level is denoted by a range. 22 The Bible. Holy Scripture. The next level: 222/224 Old Testament and 225 New Testament in general. The next level at 225 is: 226 Gospels, 227 Epistles of the apostles, and 228 Apocalypse. 229 is for Apocrypha and Pseudepigrapha of both the Old and the New Testament. These deviations of the rule: one level – one digit makes navigation in an UDC-catalogue difficult (and sometimes also the decomposition) (Loth, 1996).

The UDC is a classification built on disciplines. This means that on the higher levels we
see partitive subdivisions. First the discipline, than the subdisciplines and so on. On a lower level the objects of the (sub) discipline come and there we can have generic subdivisions. It also means that phenomena, which are the object of several (sub) disciplines, are mentioned in different classes. This is one of the reasons of the synonymy problem sketched in the last paragraph.

Hierarchical levels can be included in a description by means of upward posting, as already described above.

**Associative relations**

Associative relations can be divided into two groups: relations between terms that are subclasses of the same class and all others. The first group is denoted in the UDC by the notations. The second group is denoted by a *see also* note. In most printed versions a → is used for the associative relations. The number of associative relations varies for the different classes of the UDC. In general the number of see-also relations quite low. About 11% of all notations have one or more associative relations of the second group. Most of these *see-also’s* point to other notations in the same scientific discipline.

4. **Conclusions**

Almost all complex UDC-notations can be broken into the composing parts. Exceptions may occur in ranges and with punctuation marks (i.e. ; and +) in alphabetical additions. Of the 175,000 simple notations in the bibliographical database 8000 were truncated before a description could be added. About half of them were not existing notations. Others were hierarchical lower in a chain than the lowest in the Master Reference File of the UDC. A third group was formed by notations with a formal error as missing signs like a missing ). The last group was formed by notations which was the result of parallel subdivision but where the way to subdivide was not mentioned in the regular way.

Not always this description gives the exact meaning of the notation. Often it is necessary to add the description of a notation higher in the hierarchical chain. This always has done to notations from the main table.

The result of decomposition of the complex notations is that the search possibilities can be improved. In the first place it is possible to search on the parts. How to find all notations that contain the auxiliary of place (498)? Further a text can be added to the notations. For most users it is easier to search with words. There are some indications that searching with words in the added description even can give better results than using the classification scheme and the index. The scheme with the index gives no easy access to notations with special auxiliaries or notations, which are the result of parallel subdivision.

**References:**


