World Library and Information Congress:
69th IFLA General Conference and Council

1-9 August 2003, Berlin

Code Number: 032-E
Meeting: 126. Classification and Indexing
Simultaneous Interpretation: -

UDC implementation: from library shelves to a structured indexing language

Aida Slavic
University College London
United Kingdom

Abstract:

The UDC is attractive to different stakeholders across the information sector because of its wide-spread application, large vocabulary and availability in an electronic format. Modern information retrieval systems have the need but also the capacity to support flexible and interactive retrieval systems. The role of classification in such systems is to serve as an underlying knowledge structure that provides systematic subject organisation and thus complements the search using natural language terms. There are, however, specific requirements that must be satisfied in order to make efficient use of classification and these are not well known outside the library domain and are poorly implemented in library systems. This is especially the case for synthetic classifications, such as UDC, because its elements are meant to be manipulated by the system to fulfill different functions (a flexible systematic display, browsing or search purposes). This report summarizes the most important functionalities of the UDC that need to be taken into account during the implementation process. Important issues about the relation between the UDC schedules in electronic form - UDC Master Reference File and a classification tool (an authority file) that may be built on it, are highlighted. A better understanding of the UDC system's functionality may improve or facilitate its implementation and lower the costs of system maintenance which may be relevant for both prospective users and legacy systems.

I. Background

There are several areas of activity in the information sector at present that make it necessary to disseminate expertise in the implementation of UDC. These activities are related to both existing bibliographic systems and new users from the non-bibliographic sector. Firstly, there is a great number of libraries and bibliographic services and different legacy systems that are using UDC which do not fully exploit classification. A growing number of information gateways and union catalogues are being created that include different resource collections on a national or international level. Increasingly, users are also demanding a more efficient and more interactive information retrieval process than the majority of OPACs tend to offer. UDC data exists 'buried' in the bibliographic system in many European libraries and is not properly exploited. Furthermore, UDC can provide the necessary support in a multilingual and multi-script environment within a global information space. Also, in this environment, UDC can be used as a mapping mediator between indexing systems but this potential is mostly wasted and left unused.

In spite of a great deal of literature on UDC automation there are still many misconceptions among librarians and non-librarians about what can be achieved with classification systems such as UDC. This paper will attempt to revisit some of the well known issues in the light of common implementation scenarios based on the UDC schedules in the electronic format - the UDC Master Reference File (UDC MRF). UDC MRF is electronic form of the standard version of the UDC, owned by UDC Consortium http://www.udcc.org/mrf.htm. It is updated annually and distributed every January as ISO2709 or text files. MRF can be bought based on the annual licence agreement that can be purchased as the whole classification or, since 2003, some of its parts.

2. Implementation policy

UDC is applied to the organisation and indexing of electronic information resources, web pages, printed documents and/or realia. Irrespective of the application of UDC and irrespective of the metadata standards that are going to be chosen to carry classification data, there are some general issues that need to be tackled. The starting point in thinking through an implementation policy may be built around the following questions:

- What are the functions of subject information retrieval that need to be supported: searching and browsing; only browsing; only searching?
  If:
  a) searching and browsing: will easy transition from searching to browsing be provided?
  b) only browsing: is it going to possible to start browsing from any point in hierarchy? Will there be provision for 'see also' reference linking within hierarchies? Is classification notation going to be displayed together with class description?
  c) only searching: is an appropriate alphabetical search index going to be provided? Would it be possible to search both numbers and index terms?
- Is UDC going to be used alone or alongside some alphabetical indexing system (thesaurus or subject heading system)?
  If YES: How are these vocabularies going to be linked to the UDC: through classification authority data or through a search index only?
  If NO - An alphabetical subject index needs to be built. Is it going to be based on UDC MRF only? How it is going to be expanded, maintained? What form is the index is going to have: simple alphabetical index, chain index, relative index?

- Are there any plans to expose the collection and make it part of some larger information gateway (multilingual?) where UDC will have to be mapped to some other indexing system? Are there plans to support automatic classification in the future?
- How is it envisaged that a surrogate structure, content and syntax can support classification? Are metadata resources embedded or standalone? Which metadata standard/format will be the carrier of the UDC index and which metadata elements/fields are supporting the use of classification? What kind of format/encoding is available to hold UDC?
- How do a cataloguing/indexing policy and metadata standard relate different subject data (persons, events, coverage, topics): is it going distributed in different fields/elements, how are these fields going to be ranked and linked to form search indexes and be used by interrogating software; for which of these subjects is UDC is going to be used?
- Will subject data be supported by an authority file, and how is the metadata architecture going to relate to the document description and the authority file? Is the authority file going to be kept external to the system, or is it going to be shared by different systems or used for functions such as mapping and cross collection searches?

Some of these questions may be more relevant than others, depending upon the purpose of the system, but it is certainly worthwhile to put together a list of requirements based on the chosen policy. Most of these things are not necessarily hard to implement.

Irrespective of the choice of indexing system, there is an important but often neglected step: agreement on an indexing policy. This is not specific to classification or, indeed to UDC, and is outside the scope of this paper. However, such a guideline or document, apart from being common sense, is paramount for the success of a system and its efficiency in resource discovery. Classification schedules always leave the freedom of choice to classifiers, and this is even more the case with synthetic classification. Although the existence of a classification authority file may help support consistency and indexing control, there are still some general policy rules to be recorded. Decisions and guidelines need to be made with respect to exhaustivity and specificity in indexing. Also, things like the treatment of persons and personal names that can be added to a class mark, and places and events as a subject need to be considered. UDC can contain information that is, in MARC and other metadata formats, usually held in other metadata fields/elements such as *language of the resource*, *audience*, *external form* and *format* or *coverage*. It is necessary to decide whether repeating this within the UDC number may be useful or not.

Within the indexing policy, care needs to be taken over UDC specific issues. Often, in metadata guidelines and recommendations, indexers are led to believe that classification should be used to the highest possible level of specificity [1]. While this may well work with smaller and enumerative classifications such as the Dewey Decimal Classification (DDC), it produces cumbersome and undesirable results when applied with UDC, which is three times larger, highly synthetic, and can produce extremely indexing terms. Also one may need to record decisions in relation to the citation order in synthesized UDC numbers as this can be changed so as to produce a useful arrangement of the resources. Last but not least, if a subject alphabetical index to the classification is created, the rules for the alphabetical subject index to the classification should also be recorded. Procedures for the treatment of homonyms, synonyms, compound terms and hyperlinking of associative terms should be discussed as a part of the system design or at least anticipated as needing solution later in the process.

3. UDC implementation: functional and system requirements

The are two ways of applying UDC: a) using only simple numbers, or using pre-combined numbers as simple numbers b) using a synthetic (structured) index. Depending upon the scope and objective of the use of classification, both approaches raise issues that need to be solved by implementors. Some implementation and maintenance issues that have been mentioned are related to the way UDC data are made available in the UDC MRF, others are related to the way classification numbers are going to be used in an information retrieval system. Both aspects are addressed below. Whilst the first set of issues can be more or less alleviated by preparing a different and richer export of the classification data, the second depends on the creation of appropriate tools to manage and control the use of classification data.

3.1 Implementation of the UDC with simple, non-synthetic notation

The least complicated approach in using UDC covers both the use of simple numbers only, and the use of pre-combined numbers treated as simple numbers. The UDC standard edition, with its current set of 66,149 class numbers, can be used by choosing to deal with simple classification numbers only. These numbers can be taken from the main schedules or common auxiliary tables of the MRF and they will be detailed enough to satisfy many users. In other words, UDC can function as a straightforward taxonomy or enumerative classification. This aspect of UDC is often exploited for shelf arrangement in smaller libraries, especially in central Europe, where UDC is used in public and school libraries. Also subject gateways and portals on the Internet that deploy UDC tend to use it in this way [2]. Applied as an enumerative, non-synthetic classification, UDC serves the simple purpose of systematic browsing. When applied in this way UDC has very similar functionality to the DDC, the only difference being that UDC has a bigger and more specific vocabulary and does not contain as many enumerated, ready made compound terms as is the case with DDC.

Filing UDC with only simple numbers does not require much implementation effort. Classification notation, in this instance is a simple text consisting of numbers and meaningless punctuation (a decimal point) after every third digit. Numbers are automatically filed correctly by any computer system. More often, however, one may find UDC numbers created in a pre-combined way, but treated as simple notation. This is often the case with library systems, mostly as a result of the way MARC formats have been supporting classification data as a single string of characters only, irrespective of whether the content is a single or pre-combined, structured index term. The correct filing of these numbers is difficult and it results in a disturbed systematic order that does not follow the sequence of subjects from *broader to narrower/general to specific*, which is paramount for supporting browsing functionality. Also, this allows the search of only the first element of notation while others cannot be used for retrieval.

The use of UDC as an enumerative classification (either with simple numbers or with pre-composed numbers which are treated as simple) may well serve its main purpose if class number captions (descriptions) are added to the retrieval system so that beyond numbers, terms are available for search and are added to the systematic display at the end-user interface. The UDC MRF is a good source of index terms, that can be harvested not only in the field of caption (description) but also from the notes and examples of combinations [3].

3.1.1 Implementation issues and recommendations

Source of data. In using the UDC MRF as a source of classification data, it should be noted that not all numbers provided are simple and implementors interested in this level of UDC use, should bear this in mind [4]. In the main tables there is a small but unknown number of entries consisting of a combination of single main number and common auxiliary such as *94(680) History of South Africa*. These entries are not marked as such in a database. Also, there are numbers that are actually the combination of two main numbers or two auxiliary numbers such as span combinations in *562/569 Systematic palaeozoology* or in common auxiliary numbers for time e.g. *"321/324" Seasons*.

Extracting single numbers automatically from the UDC MRF, therefore, may not be so straightforward. Combinations of main numbers and special auxiliaries are indicated with a special field, while, as mentioned above, the combination of a main number and common auxiliary such as 94(410) is not marked for automatic processing. This is a drawback that will be corrected eventually.

New implementors, especially those providing access to information resources on the Internet, are buying the UDC MRF in order to extract single numbers or selections of them alongside their descriptions. The MRF exported for distribution to publishers and libraries is the so-called user MRF (UMRF) and it does not contain any administrative fields or even data particular to database management. Therefore, there is not enough data to extract automatically, for example, only single main numbers and no entries that are a combination of main numbers with special auxiliaries, as this information is not made available in the UMRF text file.

What would be desirable is to provide implementors with the complete MRF text data together with the MRF Manual which would provide all information on the structure and field contents [5]. The UDC Consortium should provide more choices of UDC data formats. Conversions to different MARC formats, for example, would ease the import of data to MARC-based library systems. These aspects, as well as some changes to the MRF database are currently under discussion [6][7]. There are some fields in the MRF that have never been fully used, such as the field for index terms as this was left to be added by the individual publishers of the UDC. New users of the UDC would appreciate this additional value to classification and this is another area with room for improvement to be addressed by the owners of the UDC.

Retrieval functions. Normally UDC's expressive notation allows for hierarchies to be linked to the length of notation without the need for any special adaptation for filing and display. Right truncation will lead to the broader class level which can be exploited to broaden the search. For instance, searching 004.415# will give results that include all the divisions that follow. This will also work for pre-combined numbers treated as a single string of characters. However, as pointed out by Buxton and Riesthuis right truncation does not always lead to the broader category. For instance the broader category of *563.4 Spongiaria*. Sponges is not 563 but *562 Invertebrata in general*. (Buxton, 1990, Riesthuis, 1998). This is often the case with a use of span (i.e. when a class is defined as an area covering a number of subsequent classes such as 562/569), but can occur elsewhere. This is more an exception than a rule and, although it is being gradually corrected through the revision process, it remains a feature of classification that cannot be properly managed by the simple application of the UDC without some control over hierarchies as well as the notation itself.

If implementors choose to use common auxiliary numbers (e.g. place, time, persons etc.) independently as single numbers, as well as main numbers, this will need special attention as these numbers will contain arbitrary symbols and will be automatically filed before the main numbers. Their order is going to be different from the one suggested by the UDC system. One solution to this is to enter classification data using prefixes that will serve to indicate filing order and will not appear on the display.

Implementors of classification with simple numbers only should bear in mind that the level of specificity is very much restricted in this use of the UDC and that the need to use some kind of combination of numbers may appear very early in fully faceted major classes. This is the case, for instance, at *821 Literature* and *94 History*. With the trend of present revisions moving UDC towards more faceted structure, this situation will happen more frequently. In order to make a difference between, for instance, *English literature 821.111* and *American literature* one has to use the common auxiliary of place *(73) United States of America*. Similarly, to obtain the number for history of individual countries one has to use the number for history 94 and common auxiliary for place to denote the country and if necessary the common auxiliary of time to denote the period, e.g. *94(410)"16" History of the British Isles in 17th century*. This is the reason why libraries use pre-combined numbers although they tend not to treat them as such in their system.

Management of classification. If used for browsing at the end-user interface (OPACs, portals or subject gateways) UDC class numbers should be either used alongside their descriptions or should be completely omitted, in which case the hierarchy should be displayed by indentation of the class description or via some other graphical aid. Most new implementors of the UDC tend to choose a simple application to avoid the hassle and additional problems in filing and display. However, even where a system can deal with class number sorting, as is the case with simple UDC, it may be necessary to manage the classification as separate authority data. This would allow the addition and management of all data necessary for use in combination with class numbers in order to support the functions of browsing and retrieval:

a) class number caption (description)
b) search terms which are not present in the caption and support searching and positioning in the hierarchy
c) '*see also*' references that have value for browsing
d) establishing a relative hierarchy, independent of notation in order to handle occasional notational deviation in the notational hierarchy
e) filing of common auxiliaries used as single class marks, where the symbol will be used for display purposes and not for data processing and sorting

3.2 Implementation of the UDC with synthesised, pre-coordinated notation

UDC can be distinguished from other bibliographic classifications because of its powerful synthetic feature. The advantage of synthetic classification is that it allows coverage of an unlimited number of subjects and their combinations with a limited amount of simple concepts. Synthetic features make classification more hospitable and expandable and more powerful in indexing. Synthesis, however, adds to the complexity of a classification system which then requires more knowledge of syntax rules and more support in terms of management tools. It is useful to remember that synthetic features are going to be exploited even more in future and will be further facilitated through the 'facetisation' of the UDC [8] [9]. Many classifications rely on some kind of synthesis within their schedules, mainly to economise space. UDC is, however, equipped with reliable mechanisms to support and manage unlimited synthesis at several levels:

a) among two or more main class numbers, using symbols that express the relations between two subjects
b) among main class numbers and one or more common auxiliaries
c) among main class and one or more special auxiliary numbers
d) among one or more common auxiliary numbers
e) among one or more special auxiliary numbers
f) between UDC main numbers and some other external vocabulary
g) between UDC main numbers and any other alphabetical extension used for further specification

Pre-combined numbers tend to be constructed during the process of indexing and are rarely enumerated in classification schedules. In the revision process, compound concepts are regularly cleaned out from the UDC system and are replaced with a combination of simple concepts. A UDC classification number when applied for indexing is therefore best understood as a structured, precoordinated indexing term that has its vocabulary and its syntax similar to any other pre-coordinated indexing language. The meaning of each element remains the same outside and within combinations and can be searched as in a post-coordinated manner. For instance, one will use common auxiliaries *(73) United States of America* and "18" *19th century* in an unlimited number of combinations such as: *94"18"(73) History --19th century-- USA,* or *821.111(73)"18" American literature -- 19th century* or in *321.7"18"(73) Politics -- Democracy -- 19th century -- USA.* Therefore, searching *(73)*, will retrieve every item related to the USA, and searching *"18"* will retrieve everything related to 19th century, no matter the subject.

When used to its full synthetic capabilities there are two major requirements for handling UDC: a) filing of complex numbers; b) searching of each individual element that is built in the pre-combined classification numbers. Filing of UDC simple and pre-combined numbers serves the purpose of subject presentation from general to specific. The classification system achieves this through the combination of filing rules and rules used for building a sequence within a pre-combined number. The management of the UDC therefore means the control over individual numbers whether they are used alone or built into a pre-combined number. This control should rely on formatting classification data in a way that each element of the pre-combined UDC number is recognized by the system irrespective of the symbols and facet indicators that are used for its display and irrespective of its place in a pre-combined UDC number.

These are the reasons why the use of classification and especially of the UDC, depends on the tools made available to achieve this functionality with minimum possible discomfort for cataloguers, whereby the complexity of a notation is simply handled by a system.

3.2.1 Implementation issues and recommendations

Source of data. In the process of classification the UDC MRF serves as a source of simple numbers that are used to build pre-combined indexing terms. The fact that the classification data are available in an electronic format only helps to avoid a part of the manual editing and clerical work, but the real help in document indexing is to have access to pre-combined numbers and to ensure their easy reuse. This is usually achieved through the creation of a classification database or an authority file which grows with its application. Pre-combined numbers that exist in the MRF in both the class number field and in examples of combinations, may be used as a ready-made source of indexing terms further to populate the classification authority file. These numbers do not have their structural elements encoded and some manual editing may be necessary to make them fully functional within the existing tool. The real indexing tool in this case becomes the file containing pre-combined UDC numbers built during the process of collection indexing. Easy use and management of this file is therefore paramount for the proper functioning of classification. The main goal is that once created, a pre-combined UDC heading can be linked to an unlimited number of items within a collection without much manual editing.

Depending upon the indexing and implementation policy, the total number used within a collection may be significantly smaller or significantly larger than the MRF itself. If the policy is to use the classification alongside other indexing languages, its function of gathering and aggregating would be more important and therefore the total number of pre-combined UDC classmarks may be three to four thousand even for a collection of a few hundred thousand resources. This approach is characteristic of some large public or medium academic libraries in East European countries such as Hungary, Croatia, Slovenia, etc. Libraries that use UDC as a main indexing and retrieval language with a policy to express high specificity of individual items may have hundreds or thousands of different pre-combined UDC numbers to manage. This is the case for instance in the Central University Library in Bucharest which has hundreds of thousands, or in *ETH-Bibliothek* which has around 60,000 of different pre-combined numbers [10].

A scheme of properly structured and encoded pre-combined UDC numbers can be a valuable resource in information exchange and can be shared, adapted and incrementally built upon by many implementors. This kind of reference tool provides a literary warrant for concepts that are used and may be a valuable resource in development of new vocabularies or indeed in revision of the UDC schedules themselves [11].

Supporting retrieval functions. The functional requirements for browsing and retrieval of pre-combined UDC numbers were summarized by Buxton in 1990. He stressed the following functions: the need to search UDC numbers with all the symbols that may be used; the ability to file pre-combined UDC numbers, the ability to search by truncating numbers; the ability to search separately each common auxiliary, the ability to search intermediate numbers when pre-combined UDC numbers contain a stroke (span) and the ability to truncate within a number. Buxton suggested breaking a string of pre-combined numbers at least by a space but also suggested the replacement of UDC symbols by letters following the example of AUDACIOUS system [12]. Post-coordinate searching of UDC elements and their combination with index terms is especially important for classifiers. Search expressions like 'rabbit' AND '6#', or '7#' and 'technique' with or without truncation are the best way with which one can position oneself at a certain place within the classification.

Most of these issues can be solved if the classification is implemented to deal with UDC numbers encoded pre-coordinatedly. This would allow the easy creation of rules for pre-combined UDC number sorting. Namely, in order to express the subject hierarchy, UDC numbers should be filed according to a specific set of conventions. For example, *73 Plastic Arts*; *73+75 Plastic Arts and Painting*; and *73/75* (subject covering sequence in the schedules 73 Plastic Arts, 74 Drawing, 75 Painting), need to be filed in the following sequence: *73+75* > *73/75* >*73.* This is because each symbol has its place on the scale from general to specific, and two main numbers connected with + and / give classes with broader subjects than a single class number. However, two main classes connected with : (colon), e.g. *73:75 Relationship between Plastic Arts and Painting* are always a narrower subject than a single number and needs to be filed after the number on its own e.g. *73 > 73:75*. This purely intellectual ordering should be supported by the system which will ensure the processing of these structured indices without relying on the symbols used for visual representation and display.

Another reason for having access and control to every piece of a structured UDC number is related to the flexibility in combinations of elements. Apart from the general recommendation to cite common auxiliaries in the sequence *time*, *ethnics*, *place*, *form*, *language*, the order in which UDC numbers can be combined (i.e. citation order) is flexible. Depending on the intention in presentation and arrangement some collections may want to provide different approaches for their users. In history, for instance, it is possible to have the following order: main number, time, place *94"18"(410) History - 19th century- British Isles*, which will present history by time and then by countries, while another display may allow grouping of history by countries *94(410)"18" British Isles - 19th century.* When one can establish access and processing control over each of the constituent elements, it is possible to provide different displays to suit users' preferences.

However, if UDC pre-combined numbers are kept and managed as a single strings of characters, it may still be possible to retrieve single numbers within a pre-combined string and use UDC for post-coordinated searching. This can be achieved using a specially written program that will enable the deconstruction of numbers into their constituent elements based on the algorithms extracted from the UDC syntax. Riesthuis has suggested the set of algorithms that may be used to write a program which would decompose UDC numbers [13][14].

Management of classification. As has often been emphasized, the implementation of the UDC as structured and properly encoded authority data is paramount for all the functions that classification may have in the information retrieval process. Many libraries have developed classification tools based on their own experience and needs in authority control and to date these are mostly proprietary solutions and examples of good practice. In these systems, classification is usually linked and mapped to other indexing systems or to its own subject alphabetical index. Some libraries are incrementally building thesauri or subject heading systems based on existing UDC data. New users outside the library domain expect to manage classification systems in some form of authority data and expect to find some UDC authority files available for sharing. Those implementing UDC as a synthesised and pre-combined numbers may choose one of the following approaches:

- to provide for structuring and encoding of separate UDC elements within the bibliographic description/metadata
- to maintain classification as separate data with links to an information retrieval system
- to have both the structured index in bibliographic/metadata and rich classification authority data maintained separately.

The first approach would help in sorting and retrieving UDC numbers but will fail to provide a link between classification numbers and their descriptions and search index terms, will not be able to implement 'see also' references and will not help in cataloguing as UDC numbers will always have to be re-typed and re-entered into the system. The second and the third approaches that keep classification data in separate files are much more efficient. The third one has the advantage of being more robust and reliable as UDC numbers can be properly processed and exchanged even when the authority file is detached.

Classification authority data as referred to here comprises not only a simple control file with the purpose of ensuring the access points and uniformity of UDC headings, but rather a fully functional tool that serves to hold, manage, maintain and share classification data. Its purpose would be to make unlimited use and re-application of UDC data and therefore serve as a time and effort-saving help in the classification of resources.

4 Importance of classification data formats

The higher the level of its data formality, the more powerful a classification becomes. At the same time, it become less suitable for human handling which demands more intermediation and more sophisticated mechanisms for its implementation and exploitation. This is typically the case with synthesised UDC numbers when used for pre-coordinate indexing. At present there is no accepted or suggested UDC data format that can fully cope with all the demands of handling pre-combined UDC numbers. There are, however, formats and data structure analyses available that may help in collecting the information necessary to create a fully functional UDC database which can serve functions such as classification authority control, maintenance, exchange and sharing, information retrieval and appropriate classifying tool functionalities.

The first source to be mentioned is the data structure as used in UDC MRF, which is available in the MRF file Manual, and explained in a number of articles and on the web pages of the UDC Consortium [15][16][17][18]. But the MRF data structure does not comprise all the information that needs to be accommodated in order to serve adequately all the required purposes. In particular, it lacks the structure needed to automate fully handling of the different data elements in a pre-combined UDC notation.

Another source useful to help gain an idea of what data may need to be included in a classification format is the MARC21 Format for Classification Data [19]**,** developed in 1991 and updated in 1995 by the Library of Congress, for the purpose of managing the Library of Congress and the Dewey Classification systems. But again, as these two classifications are enumerative and not synthetic, this format will not respond to the UDC needs as outlined in this paper. Notably, the fields to record a classification number (i.e. classification heading) allow only the accommodation of a simple string of characters. In other aspects, however, this format may be a useful source of what data are needed such as: caption, scope notes, instructions, examples of combinations, relative hierarchy (broader class), index terms, structure of other information such as description, index terms, see also reference, replace/replaced by, etc.

More recently, under an initiative by the Permanent UNIMARC Committee, a UNIMARC Classification Format was developed. This work started following a preliminary study - Requirements for Format for Classification Data, by E. W. Woods, 1994 – whose recommendations included the applicability for different classification schemes, for multilingual demands, for authority control functions, etc. As a result of this initiative a Concise UNIMARC Format for Classification Data [20] was made available for public discussion in 2000, and is still in a draft and unfinished form.

It was expected that this new format would pay more attention to synthetic classifications such as UDC, which basically means the support of processing and handling of UDC structural elements that are the greatest obstacle to the proper exploitation of UDC in library OPACs. As it currently stands, it offers no more than what is already made available by the MARC 21 format, thus it still does not provide for the adequate treatment of pre-combined classification headings. If these details were provided, it would fulfil the needs for multidirectional access, easy search processing and correct filing. In particular, the format does not especially address the pre-combined UDC numbers issues, notably the search of separate meaningful elements or the management of global changes in component elements. This is the major drawback of this new format. The remaining of the UDC recording needs, such as description, index term, scope note, application note, examples, *see also* references, as well as administrative management data, are covered.

Building a more sophisticated data structure to support classification features and better application functionality will certainly repay any investment. What has happened so far is that implementors in libraries that use proprietary systems, and users of UDC outside the library domain, have been in a better position than libraries with standard systems. While in the first case systems are developed according to locally defined design and data structures, libraries with standard, MARC-based systems faced reluctance and even refusal on the part of vendors to implement changes with respect to deviations from the MARC official data structures. Many libraries have attempted to negotiate with vendors to allow the splitting of UDC numbers within bibliographic data. This was discussed earlier this year by the udc-forum discussion group [21]. Some colleagues have been successful with INNOPAC but not before they turned to the USMARC maintenance body and obtained approval to use the subfield for the UDC with code 'x' to separate each element of the UDC symbol. Other colleagues were less fortunate with, for instance, the Vubis library System, with respect to similar demands.

What we may at least conclude is that none of the situations mentioned are good. The best outcome would be a standard, flexible and sufficiently complete data structure that could serve all purposes. This would present a good opportunity for the UDC Consortium to make the UDC MRF available in such a format and users could obtain both the structure and data ready for implementation. The wider metadata community has already accepted that knowledge organisation tools are better handled as independent data, external to the metadata record itself. The value of pre-coordinated languages usually becomes lost if syntax relationships are not encoded. UDC is recommended, being a standard knowledge organisation system, for use in many metadata standards such as Dublin Core or Learning Object Metadata, Encoded Archival Description etc. Most of these standards allow the encoding of the scheme from which the term is taken, the term itself, and also the use of URIs (Uniform Resource Identifiers) in case the classification data are exposed and the network is accessible by different applications. Much effort is now focused on assuring the permanency of these identifiers and allowing subject and other authority data to be shared and better exploited. This is a general trend and the expertise in authority control of bibliographic databases will certainly have a chance to be deployed in a wider environment. Considering this trend may help when making implementation decisions which initially may require more effort but would soon pay dividends.

5 Concluding remarks

It is common practice to describe and analyse UDC as a self-contained indexing language with its limitations and advantages independent of implementation constraints and bad practice. The idea that library knowledge organisation tools are ready made, off-the-shelf tools that are going to solve all the problems in resource discovery is, however, gradually being replaced by more pragmatic approaches. The issues of indexing policy, cost of training and implementation are increasingly under discussion and it is now widely accepted that the efficiency and maintenance costs of classification systems, for instance, depends not only on the availability of classification data in an electronic format, but also on the tools and retrieval system built around it.

It is also generally accepted that once produced, subject data based on any indexing system is too expensive to be wasted. Modern information systems have the technological capacity and power to use and combine different tools and techniques to complement each other in order to achieve satisfactory results. Today it is common to include the costs of mapping and bridging diverse applications, formats and data structure in any system implementation. Good systems tend to change and adapt, but also mix and match different approaches and different functions to best achieve their objectives. An extreme example of this approach is, for instance, GERHARD (German Harvest Automated Retrieval and Directory) which has used UDC MRF data, an academic library UDC authority file, web page harvester and a natural language processing program to build an automatic classification tool.

The choice of UDC should be based on its scalability, openness to expansion, re-purposing and flexibility to be complemented with other systems. In an information system, a classification best serves its purpose when implemented alongside an alphabetical index or alphabetical indexing language. Classification is a robust underlying knowledge structure that provides a semantic framework with coordinated, superordinated, subordinated and collateral relationships amongst its concepts. Therefore, it can serve as a language-independent tool for vocabulary control and as a complementary information retrieval tool to support advanced interactive subject browsing and navigation.

Classifications are highly formalized indexing languages and as such are capable of serving different purposes. This is especially the case with UDC which was not created purely for library shelf arrangement. It has a large vocabulary and underlying structure and grammar flexible enough to adjust to different applications. Classification schedules are, however, a professional tool par excellence which requires expertise and intellectual labour. This requirement can be greatly reduced if appropriate technological solutions are made available. UDC, with its pre-combined notation and structured numbering is a perfect candidate to be stored, maintained and used in a computer environment. One of the prerequisites for building computer tools that handle and exploit UDC adequately is an appropriate classification format, and this has to be created with synthetic and faceted classifications in mind. Developments in this matter are of paramount importance for the current and the prospective uses of classification.

6. Notes and references

[1] Subject data in the metadata record: recommendations and rationale: a report from the ALCTS/CCS/SAC/Subcommittee on Metadata and Subject Analysis. July 1999. http://www.govst.edu/users/gddcasey/sac/MetadataReport.html

[2] Examples of subject gateways that are applying simple UDC numbers are Social Science Information Gateway (http://www.sosig.ac.uk), and Directory of Network Resources - NISS (http://www.niss.ac.uk/subject2/new95udc.html).

[3] There are over a thousand class numbers (2285), in the MRF 2002 containing at least one or more examples of combination with their textual descriptions, while 6912 class numbers have one or more 'see also' references.

[4] The latest edition of MRF (2002) contains 48,962 single main numbers, 17,187 combinations of main numbers and special auxiliaries, and 11,138 common auxiliary numbers that consist of combinations of symbols and numbers.

[5] UDC Consortium *Master Reference File Manual*, version May 2003, is written to support maintenance of MRF CDS/ISIS database and distribution of MRF. The Manual (26 pages, PDF format) can be downloaded from (http://www.udcc.org/mrf.htm

[6] Riesthuis, G. J. A. "Some thoughts about the format of the Master Reference File database. *Extensions & Corrections to the UDC*, 23 (2001), 15-22.

[7] Riesthuis, G. J. A. "A Revised Format for the Master Reference File Database". *Extensions & Corrections to the UDC*, 24 (2002) 11-15.

[8] The present revision policy of the UDC aims at restructuring the schedules based on facet analysis. This will restrict unnecessary repetition and the number of compounds in the schedules.

[9] McIlwaine, I. C. "UDC in the twenty-first century", The future of classification, edited by Rita Marcella and Arthur Maltby, Aldershot: Gower, 2000, 93-104.

[10] Central University Library in Bucharest, Romania, uses UDC for very detailed indexing. ETH-Bibliothek in Zurich is using around 60.000 pre-combined UDC numbers to which they have attached over 400,000 corresponding subject terms in English, French and German. See: Hug, H.; Noethiger, R.. "ETHICS: an online public access catalogue at ETH-Bibliothek, Zurich". *Program*, 22, 2(1988), 133-142.

[11] German Harvest Automated Retrieval and Directory - GERHARD http://www.gerhard.de, has used data from ETH-Bibliothek's authority file, in total 60.000 pre-combined UDC numbers, which with it subject index amounted to 500.000 lines of text. See: Möller, G. etc. "Automatic classification of the World Wide Web using Universal Decimal Classification", 23rd International Online Information meeting, London, 7-9 December 1999: proceedings. Editor Brian McKenna. Oxford: Learned Information Europe, 1999. 231-237.

[12] Buxton, A. B. "Computer searching of UDC numbers", *Journal of Documentation*, 46 (3) 1990, 193-217.

[13] Riesthuis, G. J. A. "Decomposition of complex UDC notation", Knowledge organization for information retrieval: proceedings of the Sixth International Study Conference on Classification Research, London, 16-18 June 1997. The Hague: International Federation for Information and Documentation (FID), 1997, (FID 716), 139-143.

[14] Riesthuis, G. J. A. "Decomposition of UDC-numbers and the text of the UDC Master Reference File", Structures and relations in knowledge organization: proceedings of the Fifth International ISKO Conference, Lille, 25-29 August 1998. eds. W. Mustafa el Hadi, J. Maniez, S. Pollitt. Würzburg: ERGON Verlag, 1998, (Advances in knowledge organization 6), 221-228.

[15] Strachan, D. "UDC revision work in FID", The UDC: essays for a new decade, edited by Alan Gilchrist and David Strachan. London: ASLIB, 1990, 1-10.

[16] Strachan, D.; Oomes, F. M. H. "The UDC Master Reference File (MRF)", *Extensions & Corrections to the UDC*, 15 (1993). The Hague : UDC Consortium, 1993, 19-28.

[17] Strachan, D.; Oomes, F. M. H. "Universal Decimal Classification update", *Cataloging and Classification Quarterly*, 19(3-4) 1995, 119-131.

[18] Riesthuis, G. J. A. The UDC Master Reference File. 64th IFLA General Conference, August 16 - August 21, 1998. http://www.ifla.org/IV/ifla64/157-158e.htm

[19] MARC 21 Concise Format for Classification Data is available at http://www.loc.gov/marc/classification/eccdhome.html

[20] Concise UNIMARC Format for Classification Data is available at http://www.ifla.org/VI/3/p1996-1/concise.htm]

[21] See the thread 'computers handling UDC numbers' at the udc-forum archive at http://www.jiscmail.ac.uk/lists/udc-forum.html, February 2003.