European Commission
Directorate General XIII

Electronic publishing and libraries

Telematics for Libraries

UNIMARC WORKSHOP

Proceedings of the Workshop

held in

Luxembourg on 13 September 1996

CEE: XIII/61

December 1996

Contents

WORKSHOP ON UNIMARC AND EU PROJECTS

Luxembourg, Friday, 13 September 1996

REPORT OF THE WORKSHOP

1. Introduction

The workshop was arranged in order to bring together representatives of various organisations and projects directly or indirectly concerned with the UNIMARC format. Its purpose was to assess progress made in removing format incompatibilities as a barrier to record exchange; to identify actions needed to sustain and continue this process, if necessary, and to discuss how to prevent similar format barriers from inhibiting future exchange of extended bibliographic information and the related electronic documents.

The specific objectives were, through exchanging information on the results of the projects to date, to:

- identify problems which have been resolved and to discuss impact and take-up of the solutions proposed
- identify the impact on the format
- discuss remaining problem areas, together with possible remedies
- identify how to take actions forward into the electronic document environment.

The programme for the day and the list of participants are given in Annex 1.

Setting the context for the workshop, the Commission referred to the meeting held in Florence in 1991 at which the findings of the UNIMARC-EC study of 1990-91 were presented. That survey had revealed much theoretical interest in UNIMARC but relatively little practical experience with the format. Five years on, UNIMARC is far better established and more highly regarded, having been adopted for several high-profile co-operative projects and as the national or external exchange format by more countries, and applied to a wider range of materials.

2. Background and context of Commission actions

The key problem of the availability – or, more accurately, non-availability – of European bibliographic records was identified in the early days of the Action Plan for Libraries. The strategy to address this was two-fold: the development of national bibliographic services, and the retrospective conversion of the catalogues of important collections. Across Europe, existing provision and work in hand was very fragmented and uneven. It was recognised that UNIMARC had the potential to overcome some of the problems of exchanging bibliographic records. Four preparatory national projects were set up to improve national bibliographic services; some of these aimed at improving the use and re-use of records in UNIMARC. A number of early projects also tackled conversions between UNIMARC and national formats: in the context of a project with the National Library of Ireland, Trinity College Dublin developed a conversion routine from UKMARC to UNIMARC;

EROMM (European Register of Microform Masters) required UKMARC-to-UNIMARC and INTERMARC-to-UNIMARC conversions, and *CDBIB* (National Libraries Project on CD-ROM) added conversions to UNIMARC from both danMARC and PICAMARC for its pilot disk "The Explorers" containing records from the national bibliographies of Denmark, the Netherlands, Italy and Portugal.

Parallel with this work, DGXIII/E-3 (the Libraries unit) launched a study into the use of UNIMARC, with the particular objective of establishing whether it was feasible to use it as an exchange format and what the problems would be. The aim was to validate promoting its use in projects as the exchange format of preference. The study (covered in more detail in the Workshop background document - see Annex 1) found wide disparities in computers, formats and scale of operation, general agreement on the desirability of measures which would simplify and make cost-effective the exchange of bibliographic records, relatively little use made of UNIMARC for this purpose at that time, and almost total non-use (and non-awareness) of UNIMARC by the book trade.

In the Libraries workprogramme under the Third Framework Programme for RTD, the goal of improving national bibliographic services was retained, but with a focus on the development of standard and internationally applicable tools, rather than on particular bibliographic resources (Action Line 1, Part 1 of the work programme). However, the calls for proposals resulted in only two relevant projects, *UseMARCON* (described below), and *HELEN* which was concerned with transliteration problems as a barrier to the exchange of information, in this instance between the Greek and Latin alphabets. Some other projects had minor UNIMARC elements in them.

Several core areas in Action Line 1, therefore, remained unaddressed. To remedy this, CoBRA (Computerised Bibliographic Record Actions) was set up at the end of 1993 under the aegis of the Conference of European National Librarians (CENL) with funding from the Commission. CoBRA provides a forum for bringing together national libraries in Europe to identify issues of common concern and the strategies for tackling them. The result is the cluster of investigative projects and feasibility studies, focusing on activities which could improve the exchange and use of national bibliographical records and services. The projects of most immediate importance to the Workshop are *UNIMARC* and *AUTHOR* (both described below). Other projects with implications for UNIMARC are *METRIC*, a feasibility study of the use bibliometric data in national databases, *FLEX*, which deals with standards for the labelling of files and records, and especially *CHASE*, which is concerned with the possibility of migrating to the UNICODE character set.

CoBRA also incubated BIBLINK, a project retained under the Fourth Framework Programme. BIBLINK aims to establish an electronic link between national bibliographic agencies and publishers of electronic material, in order to create authoritative bibliographic information that will benefit both sectors. It will investigate how the data can be incorporated in the electronic publications and extracted for use in MARC-based national bibliographic services. There are many other projects in hand whose primary objective is not bibliographic record exchange but where differences between formats materially affect the usability of the systems they are developing.

The present position may be summarised as

- several format conversions have been made: but mainly one-to-one;
- there is evident duplication of effort (for example, in conversions for the same pairs of formats), despite publication of results and sharing know-how;
- there are technical problems (for example, character sets), affecting all MARC formats where work is still needed; and
- there are some projects which have worked, or are working, in an SGML, rather than MARC, environment.

It is evident that more attention needs to be paid to

- the exploitation and take-up of research results; and
- avoiding duplication of effort, or "starting from square one" each time.

Areas for possible continuing work are:

- authority records;
- the relationship between bibliographic formats and document formats;
- records for electronic documents; and
- linking and navigating between bibliographic data (or other metadata) and the documents and resources described.

3. Presentations of ongoing technical projects

The workshop focused on ongoing technical projects funded by the European Commission and on other relevant projects and initiatives, as well as on exploring the implications of the emerging electronic document environment and its standards. Fuller details of projects are provided in the background paper (included in annex I) *Workshop on UNIMARC and EU projects : a synthesis of projects.*

3.1 *UseMARCON (User-controlled Generic MARC Converter):* **Drs Trudi Noordermeer (Koninklijke Bibliotheek)**

The National Libraries Project on CD-ROM *(CDBIB)* provided valuable experience in converting two national formats to UNIMARC and combining these with two varieties of UNIMARC on the same disk. One of the lessons learnt from the project and the research that went into it was that a generalised approach to format conversion was desirable instead of one-to-one conversions. The project *UseMARCON* is a very ambitious attempt to solve this problem.

About 50 MARC formats and their specialised variants (some of them now obsolete) have been identified. *UseMARCON* is designed as a toolbox which can be used to convert between any pair of formats which conform to ISO 2709 – PICAPLUS and MAB are thereby excluded – including variants of the same format. It was originally intended that the toolbox should be easily used by any chief cataloguer with a good knowledge of the source and target formats, but it is now recognised that to make full use of it considerable experience of systems analysis (if not actual computer programming) would also be needed.

It was decided to use UNIMARC as the central reference format with the conversions being made from source format into UNIMARC and from UNIMARC into the target format: the user would only see the input and output. (UNIMARC might itself be the source or target format). It is not disputed that UNIMARC has disadvantages in this role, but the alternatives were found to have as many or more. The idea of loading every known and conceivable data element into the core was also discarded fairly early on.

Work is now in the third and final phase of beta testing by partner and other libraries and the production of documentation; the project is due to complete in January 1997.

The deliverables consist of conversion software, including coded data tables, format descriptions and conversion rules, and documentation. The tables list data (for example, country codes) as used by each format; the format descriptions are highly formalised descriptions of the formats, each containing information about valid indicators and subfields, repeats, etc.; and the conversion rules are the formal, logical statements used by the software to govern the conversions between the pairs of formats.

The existing software package has ready-made conversions for standard UKMARC, USMARC, UNIMARC and INTERMARC, and has also been tested on the UKMARC and UNIMARC Authorities formats. To create further conversion routines for other formats or variants, the rules can be edited in real time from within the software package by the cataloguer/systems analyst, but the tables and format descriptions are not directly accessible and have to be downloaded as ASCII files and edited by a standard word processing package. Thus to convert from, say, danMARC to LIBRIS it is necessary to create format descriptions and tables for both and then edit the rules files to suit the danMARC-to-UNIMARC and UNIMARC-to-LIBRIS pairings. Obviously, the closer these formats are to one or other of the formats for which the rules and tables have been produced already, the fewer the changes required, but this work needs to be done with great care. The software has been designed to allow data which is present in source and target formats but which has no equivalent in UNIMARC to be retained.

The software has been produced in versions for MS Windows (3.1x and 95) and also the UNIX Motif environments, the latter running under the Sunsoft Solaris operating system. A Windows NT (32-bit) version is under consideration; this would have possibilities for implementation of UNICODE.

A key problem identified in the course of the project was the lack of format descriptions that are publicly available; many are in little-known languages, or out of date. Maintenance of the conversion tables and of the conversion rules is a critical issue for future exploitation of the product. Other exploitation-related problems which remain to be addressed concern marketing, distribution and support (e.g. help desks, training, demonstration).

Discussion: the main points to emerge concerned the potential of the tool and format availability.

Potential of the tool. Though UseMARCON represents a relatively modest investment in financial terms, it is a flagship project for the libraries sector. Its modular structure was recognised as one of the strengths of the software, allowing

considerable and unforeseen flexibility in its use and integration with other applications (eg for character set conversions).

Format availability. Most formats are maintained but this maintenance and its documentation is often directed at internal use. One factor in the equation is the organisational structures: some are very democratic (which may lead to long update cycles); others are dictatorial, though over-hasty revisions were also deemed inadvisable. However, since most formats are the responsibilities of national libraries, GABRIEL presents an obvious opportunity for providing information about the current status of format descriptions and where to obtain them and, ideally, making them available electronically.

3.2 UNIMARC: Dr Claudia Fabian (Bayerische Staatsbibliothek)

The project *UNIMARC* was an EC-funded CoBRA study concerning the "Feasibility of the application of UNIMARC to multinational databases"[1] for which the Bayerische Staatsbibliothek had the overall co-ordinating responsibility. For the purposes of the study, the database of files in the UNIMARC format being built up by the Consortium of European Research Libraries (CERL) provided a large and varied ready-made resource.

The database comprises records of books of the hand-press era, 1450-1830 (known as the HPB database), designed to be a tool for both cataloguers and researchers. To make a coherent database, it was decided at the beginning that UNIMARC should be used for it, records being either created in that format or converted to it. The records thus represent four centuries of printing and publishing in many European languages, in many editions and variants, catalogued according to a variety of rules and traditions, sometimes over more than a century, and using several different computer formats. Some of the machine-readable records were created book-in-hand, but many are retroconversions.

The aims of the project were to study the problems arising from differing interpretations of the options available in UNIMARC when merging records from multiple sources; the problems associated with holding, indexing and retrieving merged data from multilingual and multicultural sources; and the applicability of an agreed minimum record to such a merged database.

The project was able to take into account the availability of more than 250,000 records for early books in files from six national sources (Croatia, France, Germany, Italy, Portugal and Sweden) destined for inclusion in the HPB database. The German file (from Munich) is converted from the MAB format, while the Swedish file has been converted from the machine-readable version of a detailed printed bibliography of eighteenth-century imprints. The other four files use UNIMARC as an international exchange format, two of them (Croatia and Portugal) also as their national format.

The analysis of the files used two complementary and interacting approaches. The first was an intellectual analysis of sample records, comparing their bibliographic content and application of cataloguing rules and the manner in which these have been translated into UNIMARC. The second approach was a statistical analysis of

---

1 The report is to be published by K.G. Saur in the UBCIM publications, new series

the UNIMARC files. For this a software package was written and progressively refined which provides an analysis of the use of fields and subfields (number of occurrences, and maximum, minimum and average lengths) and also a detailed statistical breakdown and overview of the characters used in the files. The results of this analytical tool are presented as a series of spreadsheets for each file and a cross-comparison for all six files. The software can easily detect errors in the files (for example, invalid characters in the character set used, invalid subfields, miscalculated record labels, etc.), but in addition to this invaluable practical feature, it can also point to areas where further investigation seems to be necessary, because divergences and differences in application of the format become evident. These may have implications for storage, indexing and retrieval; awareness of them may also lead to agreement of common standards. The software will be made available as shareware.

The following are some significant findings:

1. UNIMARC has proved hospitable to conversions from all kinds of source formats, including those which do not recognise ISBD principles or even conform to ISO 2709 structure. The provision of 166 fields and allowance for further locally defined fields make possible very detailed specification – and also need careful monitoring in practice. CERL has found that only two groups of local fields are still necessary: alternative forms of names (79x) as a stopgap until an effective name authority structure can be implemented, and holdings data (899) until the PUC produces a definitive holdings format. Some other fields, such as Fingerprint, and Title in modern spelling, were proposed to the PUC and have now been included in the format (012 and 518).

2. Although UNIMARC is a very detailed format, the software analysis showed that only 75 fields have been used in the files examined. (Several UNIMARC fields, however, are applicable only to specific kinds of materials not represented in the HPB database). The maximum used is 50 (Croatia) and the minimum 20 (Sweden), with the average 35 for the rest. Croatia carried out book-in-hand cataloguing of about 2000 items, applying the full UNIMARC specification for antiquarian material, while Sweden's 49 000-item eighteenth century bibliography is very detailed in its content but very broad in its structure.

   The smaller or more specialised the file, the more cataloguers and/or formats tend to go into detail. Detailed format definitions are more time-consuming, and may lead to more mistakes.

   Detailed specification can permit more precise indexing and retrieval. While true in a local environment, this is a questionable advantage in cooperative databases, if the same possibilities for retrieval are not present in all files.

3. A similar consideration applies to data exchange: those using detailed formats have to carry out much expansion and retagging if they download broadly defined records, whereas fields and subfields in detailed records can more easily be cut down or merged for use in a broader system.

   Statistical analysis suggests that once the "technical" fields (001, 100 etc.) are discounted, the minimal record is very minimal indeed. Those working in

UNIMARC as a native format tend to use more fields; converted records are less specific.

Key problems identified in UNIMARC were

1. Multivolume works remain a problem; more guidance and an agreement to reduce options is desirable.

2. Coded data is very unevenly applied. It is potentially very valuable, being language-independent, unlike notes fields. Better definitions of codes and recommendations for their use are needed.

3. Character sets are a major problem, not only in ensuring that all characters can be held and displayed, but also in filing, indexing and retrieval. CERL has required double indexing of several characters.

4. The non-sorting (non-filing) characters create unnecessary problems and are used very inconsistently across the files. This may be more a question of cataloguing rules rather than formats.

Standardisation and agreement on common practice may be seen to be desirable and useful, but will be hard to achieve; moreover, existing practices often have a sound basis and should not be lightly discarded. Valuable cultural differences must not be lost. Many of the disadvantages can be overcome by better authority control.

3.3 AUTHOR: Mmes Françoise Bourdon & Sonia Zillhardt (Bibliothèque nationale de France)

The project *AUTHOR* is another activity under the CoBRA umbrella, being a feasibility study into the networking of national name authority files.

There are large national and international pools of bibliographic records bearing name access points which are increasingly controlled by automated authority files. The UBC ideal is for each national bibliographic agency to establish the authoritative forms of names for its own country's authors and organisations, and make use of the work of other agencies for foreign items. In practice the problems are (a) that not every country has an authority file, and (b) that even where these exist, they are not easy to consult.

The national libraries of France, Belgium, Portugal, Spain and the UK are cooperating in the work with the following objectives:

- to establish the technical feasibility of
  - conversion of authority files to the UNIMARC/Authorities (UNIMARC/A) format;
  - access to each other's authority files by a common test bed, and defining a target technical architecture for this; and
  - re-using authority data in current cataloguing;

- to achieve the following results
  - the creation of re-usable conversion tables from national formats to UNIMARC/A;
  - the identification of problems encountered and the submission of recommendations to the PUC for improvements to UNIMARC/A;

- access to authority data via Z39.50 and the World Wide Web;
- the elaboration of proposals for the minimal content of an authority record, in co-operation with the IFLA UBCIM working group on this topic established in May 1995; and
- the design of a target technical architecture accessible to other libraries from the test bed platform.

Given the work of the libraries involved, *AUTHOR* has to deal with four sets of cataloguing rules, five cataloguing languages, four MARC formats and four different hardware/software environments. The partners' examination of UNIMARC/A and national formats has shown that while conversion to UNIMARC/A should prove relatively straightforward, conversion back again would at the moment be very difficult, if not impossible, and has revealed a number of deficiencies in UNIMARC/A which should be redressed. More coded data in fixed fields instead of notes, where possible, would help to improve the international nature of the format. Great care will need to be used in the conversions: similar data elements coded variously in the national formats need to be mapped to the same UNIMARC/A field or subfield, and conversely, data of different types not properly distinguished in the national formats need to be mapped to their correct individual UNIMARC/A fields or subfields.

The partners have defined their needs as being:

- the ability to search on-line (rather than from CD-ROM), so that the data are up to date;
- the display of records in UNIMARC/A format;
- re-use of retrieved data by copying and re-keying, not automatic downloading, until UNIMARC/A-to-national format conversion tables are shown to be feasible and effective.

A prototype server will be built to simulate and test access to the files of authority records supplied by each of the partners. It will make use of the work of two other European projects, *UseMARCON* for conversion from national formats to UNIMARC/A (and possibly vice versa) and *EUROPAGATE* which has developed portable software providing a Web gateway to Z39.50 servers. The AUTHOR prototype will test the feasibility of eventually establishing a distributed network with records being converted to UNIMARC/A on the fly.

3.4 Overview of other relevant projects

ONE (OPAC Network in Europe): The purpose of the project is to establish a service infrastructure for searching library catalogues in Europe which can be extended to include resources world-wide through the Internet, and can be further expanded to allow ordering of publications found through searching. The project will define the functional requirements for an OPAC network in an European context. It will also establish a trial service between the users and the database (catalogue) providers participating in the project. International standards for catalogue access, ISO/SR and Z39.50 will be implemented in different technical environments. A set of software tools, intended to be portable to a wide range of system platforms, will be developed. These tools will provide additional functions such as conversion between different MARC formats for bibliographic records and character set conversion.

The following conversion tables are being developed for ONE's on-line converter: local MARC to/from UNIMARC; local MARC to/from USMARC; and local MARC to/from UKMARC. Phase 3, the first practical test phase, exchanged records in USMARC format. In the trial service resulting from the project, USMARC will be only one of several possible formats to convert to and from.

CHASE (CHAracter SEt standardisation – migration strategies to UNICODE): The principal aim of CHASE is to encourage the adoption and implementation of UNICODE by national bibliographic services, by establishing both the feasibility of using UNICODE as an exchange medium and also as the encoding standard in the source systems. Work to date has developed a series of conversion routines from the character sets in the national bibliographic files of the libraries involved to UNICODE. Work is currently ongoing on testing the results for record exchange purposes. These were discussed fully in a 2-day end of project Workshop in late November 1996.

***KSYSERROR (Knowledge-based system for consistency in bibliographic databases)* (now renamed *DELICAT*):** Cataloguers devote considerable resources to cleaning up records. *DELICAT* is a project to develop a generic tool for detecting such errors in bibliographic records, starting with a survey of the kind of errors found in national bibliographic files. *DELICAT* is designed to work through a client/server link or any network connection. Originally, the samples were to be only UNIMARC records, but now multiple MARC formats will be tested. The project will exploit UseMARCON and its format checking tables. The pilot version will be run off-line, with examination of whole files; it is hoped that this will be ready by early 1977.

BIBLINK (Linking publishers and national bibliographic services): The main object of the project is to improve national bibliographic services through better links between publishers and national libraries or bibliographic agencies. BIBLINK has been divided into two distinct phases, each expected to last about eighteen months In the first phase the scope of the project has been defined more precisely and information is currently being collected on metadata formats, on methods of uniquely identifying electronic publications, and on data transmission methods between publishers and national bibliographic agencies. Next work will investigate the authentication of publications and corresponding metadata. Considerable attention is also being given to consensus building with publishers. In Phase 2 of the project, the prototype demonstration system will be developed and installed at the sites of the project partners and the participating publishers for trials.

4. International developments : the initiatives of IFLA and OCLC

4.1 IFLA UBCIM's Permanent UNIMARC Committee: Mme Marie-France Plassard

The projects sponsored and funded by the EC have been of great value to the PUC. Their findings and suggestions have directly or indirectly resulted in improvements and extensions to the UNIMARC format and guides to its use. The recently published *Guidelines for using UNIMARC for older monographic publications (Antiquarian)* (Guideline no.3, 1996) is a good example.

Apart from the continuing process of amendment of the format, the PUC has a number of other important issues on its agenda, some involving other IFLA

Divisions and Sections and having implications for UNIMARC even if not involving immediate changes to the format:

- A classification format for UNIMARC – the options are (a) adapting the USMARC Classification format, (b) extending UNIMARC/A or (c) developing a UNIMARC Classification format;
- Document Type Description (DTD) for UNIMARC (arising from a recommendation at the ELAG meeting in Berlin, April 1996);
- *Functional requirements for bibliographic records** – this major study was issued in May 1996 as a draft report for world-wide review by November 1996; after the final results of this review have been received, minimal level records will be considered again;

  * http://www.nlc-bnc.ca/ifla/VII/s13/frbr/

- IFLA Working Group on Minimal Level Record and the ISADN (International Standard Authority Data Number) – this would probably entail improvements to UNIMARC/A.

Main problems and focus for the discussion were the infrequent meetings and tight budgets limiting what can be achieved. The PUC is keen to promote UNIMARC and welcomes users' interest, questions and suggestions. Questions and proposals for the PUC should be sent to Mme Plassard, who is its Secretary, but it should be borne in mind that the Committee meets only once a year (around March/May), so papers for consideration should be received well before then.

4.2 OCLC Online Computer Library Center, Inc. : Ms Janet Mitchell (OCLC Europe)

In February 1995 the OCLC Board of Trustees approved a number of product enhancements in order to support OCLC's international growth, one of these being the development of a UNIMARC capability . In the same month OCLC made an agreement to load the Czech National Bibliography into the OCLC Online Union Catalogue (OLUC); this agreement specified that the records should be delivered in UNIMARC format.

It should be emphasised that this development is *not* a research project but the creation of a production facility.

OCLC has considerable experience of format conversion. As the OLUC has spread its net further afield there has been an increasing requirement to import and export records in formats other than USMARC. Up to now OCLC has made use of conversion software from third parties. This has led to a proliferation that is becoming increasingly difficult to manage, not least with regard to the problem of ensuring that changes to the USMARC format are taken into account and the conversion programs modified accordingly.

An in-house facility which would enable OCLC to exchange records in a standard format with a wider range of overseas customers, and which would be under OCLC's own control, was therefore considered to be a necessary development. UNIMARC was seen to have good documentation and organization for maintenance, to be a format adopted by many national libraries for exchange purposes, and to be the chosen format by many libraries in Central and Eastern Europe.

A human problem, encountered early on, was that all OCLC's development experts are in the USA and are familiar with USMARC. Accordingly, the OCLC UNIMARC group met representatives of the Czech, Russian and Croatian national libraries in April 1996 to discuss questions such as the tolerable level of data loss, the amount of variation in the application of UNIMARC likely to be encountered, how to link bibliographic records and how to link bibliographic to authority records – something USMARC either cannot do, or does in a more rudimentary way – and what effect the different cataloguing rules may have.

Systematic comparison of the formats revealed great differences between them – character sets, treatment of main entry, embedded fields, etc. – and the conversion specifications have to be made subfield by subfield, both ways : there are no short cuts. The drafts were prepared between June 1995 and January 1996, followed by testing of the UNIMARC-to-USMARC conversion which was largely complete by July 1996, when a converted file was sent to Prague. Correction and refinement is now underway. The USMARC-to-UNIMARC conversion should be completed by June 1997.

Test files from Die Deutsche Bibliothek, ICCU (Rome), and the National Library of Portugal were also invaluable for trying out the conversion software and in revealing many of the variations in UNIMARC practice which are likely to be encountered.

4.3 Key issues and problem areas

There are in practice currently two major international formats - UNIMARC and USMARC - with the balance between the two likely to be affected by the proposed harmonisation of USMARC, UKMARC and CANMARC. A major factor in the choice of formats by libraries is the predominance in the library systems market of US-originating suppliers and systems which have USMARC as the default format (or the import/export format) for their databases. This pushes many libraries, especially those automating for the first time, to adopt USMARC. However, there are systems, including North American ones, which offer support for UNIMARC. What is needed is more exchange of information and experience from users of UNIMARC and also more pressure on and response from suppliers to support the different formats which reflect different cultures. One positive suggestion was for a coalition of UNIMARC users.

OCLC's initative in providing data in UNIMARC was applauded and the suggestion that OCLC become a Corresponding Member of PUC was welcomed by Ms Mitchell; Mme Plassard promised to put the suggestion to PUC.

5. The prevention of the development of format barriers (especially in extended bibliographic information in the electronic document environment)

Mme Catherine Lupovici (Jouve Systèmes d'Information, FR)

This session set out to examine the respective approaches to bibliographic description in the MARC/cataloguing environment and in the SGML/HTML document environment where certain bibliographic data are embedded in the document itself.

The components of the bibliographic records were identified as:

- Coded information;
- Identification of the document, including description (with notes, etc., added); and
- Access points.

Additional links are needed if the user wishes to navigate from one record to another (eg in multi-volume works) or from the record to the document itself.

The Standard Generalized Markup Language (SGML) and its variants, for example, Hypertext Markup Language (HTML), have introduced a different approach and a new dimension to the handling of documents and their descriptions in electronic form, as well as to hypertext linking and navigation. Documents are tagged according to a Document Type Description (DTD) which defines the elements in the tagged document and their relationship. Initially used to mark up authors' output so that it could be readily "translated" into formatting and style by publishers, mark-up languages are now used more widely, and their potential for indexing, retrieval and reformatting is recognised. DTDs are becoming more standardised, especially for general work. ISO 12083-1994, Marking electronic manuscripts, contains standard DTDs for books, serials, articles and mathematical formulae.

It is now necessary to think about the kind of bibliographic information in which we should invest in the future. In cataloguing, secondary data elements are used – the title, subtitle, etc. Much of this information can be derived directly if the document description is electronically tagged. Will MARC formats still be required, or some more general software for the new environment of electronic publications? These may call for reconsideration and redefinition of our concepts of titles, access points, and an identification of what can be derived automatically and of what elements constitute added value (eg authority forms).

SGML can be used to define MARC formats. For example, the University of California at Berkeley has produced an SGML-tagged version of the USMARC format, which was created with the particular requirements of Greek and Cyrillic documents in mind.

In discussion, some scepticism was expressed both about the volume of documents available in SGML and about standardisation of SGML e.g. the number of variants which suggested a situation not unlike that of MARC.

6. Key issues, problems and recommendations arising from the Workshop

6.1 Formats

UNIMARC format

- It would be useful if the work of the PUC were strengthened (with appropriate funding) to allow more than one meeting a year and subcommittees to meet and work effectively on specific developments.
- The format needs development for the recording of holdings data.

- The format should be examined to see if some simplification can be achieved where there are alternatives, and more guidance given, notably for the treatment of multivolume works and in the use of linking mechanisms.
- Linking mechanisms (both those which link bibliographic records and those which link bibliographic to authority records) must *not* be dropped: UNIMARC is more advanced than some other formats (for example, USMARC) in this area. However, conversion between formats which use these linking structures and those which do not can be a major problem.
- UNIMARC/Authorities format should be revised and developed as a high priority.
- UNIMARC must be boldly publicised and marketed. Its use in both projects and "real life" applications must be made widely known. European libraries adopting UNIMARC as national or international exchange format should band together to influence suppliers of library systems to incorporate a UNIMARC capability as a matter of course.

Conversion

Exchange of records between different formats requires conversion routines. There have been far too many one-to-one conversion routines written, often duplicating each other for the same pairs of formats (for example, UKMARC to USMARC).

- Duplication of effort must be avoided as far as possible. The further use and exploitation of *UseMARCON* points a way forward. It must be fully tested and widely applied and its use in an online environment investigated, including in Z39.50 interfaces.

Maintenance and publication

If efficient conversion is to be and remain possible, national formats must

- be regularly revised;
- promptly documented;
- disseminated and made generally available; and
- preferably published in English as well as the local language, especially if the latter is not widely known.

This should be the responsibility of the national library or national bibliographic agency, with the Consortium of European National Libraries having coordinating responsibility.

- Information about the formats should be given on the GABRIEL web pages.

6.2 Authority control

Authority control is seen as a high-value component in information storage, indexing and retrieval systems, whether MARC-based or other. It provides access and links to forms of names of all kinds in ways which cannot be derived from any one document (nor even adequately, very often, from several related documents). Authority forms add value to descriptive bibliographic data inherent in electronic documents.

- High priority should be accorded to the work of the *AUTHOR* project, and also to the improvement of the UNIMARC/Authorities format.

6.3 Character sets

There are a number of problems associated with character sets and MARC formats which remain to be tackled, including:

- Various formats prescribe the use of different character sets. Apart from individual characters, the sets may omit whole alphabets (for example, Greek).
- Characters which cannot be converted require special processing to store, represent and display them when used in another environment.
- ISO 646 (IRV), Basic Latin Set, is the default character set for UNIMARC, and is mandatory for control characters, indicators, subfield codes and coded values. This may be a barrier to the use of UNIMARC by librarians in countries which do not use the Roman alphabet.

However, projects are beginning to yield results in some areas and should be exploited further:

- *UseMARCON* (and some SR/Z39.50 projects) are providing character set conversions : this work should be coordinated and further developed.
- Progress towards UNICODE must be maintained, and the implications for amendments to the UNIMARC and other MARC formats examined, considering also the costs of converting existing systems and records.

6.3 Electronic publications

Publications in electronic form may not yet be quantitatively or qualitatively the predominant and most important, but their numbers and significance are increasing rapidly, both with original items published as electronic documents and the digitization of previously issued print publications.

It is necessary to reconsider the links between documents and their descriptions (metadata).

- Is the traditional bibliographic description appropriate?
- Is it adequate, and if not, what enhancements need to be made?
- What is a "document" or "publication" in this environment?

Anthony G. Curwen
Aberystwyth, Wales, UK
October 1996

ANNEX I

1. Synthesis of projects

2. Agenda of the meeting

3. List of participants

WORKSHOP

ON

UNIMARC AND EU PROJECTS

A SYNTHESIS OF PROJECTS

LUXEMBOURG

WAGNER BUILDING

ROOM GLESENER A

FRIDAY 13.09.1996

Anthony G. Curwen

Aberystwyth, August 1996

UNIMARC - A SYNTHESIS OF PROJECTS

1. Introduction

"The primary purpose of UNIMARC is to facilitate the international exchange of data in machine- readable form between national bibliographic agencies": this statement still takes pride of place in the *UNIMARC Manual*. IFLA published UNIMARC with the intention that it should be an intermediate format to obviate the need for conversion programs between every possible pairing of MARC formats.

The original emphasis was on books and serials, but later developments resulted in

- the use of UNIMARC for other materials, and
- the adoption of UNIMARC as a national or local format,

so that the *Manual* now goes on to say "UNIMARC may also be used as a model for the development of new machine-readable bibliographic formats". (Might this be construed as encouraging the creation of even more variant formats, which UNIMARC was designed to limit or render unnecessary?). Today UNIMARC is in widespread and growing use for its original purpose:

- through the provision of records additionally in UNIMARC format by national agencies;
- as a national format (as in Portugal and Croatia) and as the basis of others;
- as a "hidden switching language" in the UseMARCON project;
- and as the preferred format for co-operative ventures (for example, EROMM, CERL).

The purpose of the workshop is to assess progress made in removing format incompatibilities as a barrier to record exchange; to identify actions needed to sustain and continue this process, if necessary, and to discuss how to prevent similar format barriers from inhibiting future exchange of extended bibliographic information and the related electronic documents.

The specific objectives are, through exchanging information on the results of the projects to date, to:

- identify problems which have been resolved and to discuss impact and take-up of the solutions proposed
- identify the impact on the format
- discuss remaining problem areas, together with possible remedies
- identify how to take actions forward into the electronic document environment

2 Background and Context for the Meeting

In the context of the Libraries Programme under the Third Framework Programme, a number of projects and actions have been funded which set out to tackle practical issues surrounding the bibliographic record exchange between different formats. A particular emphasis was placed on the application of UNIMARC as a common format in the exchange process. These projects form a natural "cluster": in addition to this core, there are other projects which are tackling related problems, such as character sets and the relationship of bibliographic formats to document formats. Furthermore, there are other projects which, while not directly addressing bibliographic formats for exchange, are affected adversely by format differences - this is particularly the case with implementations of Z39.50.

3. Libraries Programme: Preparatory Actions

3.1 The UNIMARC Study

At a workshop on national bibliographic services in the EC, held in Luxembourg in February 1990, UNIMARC was proposed as the common exchange format for the national bibliographic agencies in the Community. Subsequently Die Deutsche Bibliothek conducted, for the European Commission, a *Study to establish the feasibility of using UNIMARC amongst EC national libraries, bibliographic utilities and the booktrade based upon their present computer facilities.* This investigated:

- the actual and potential use of UNIMARC, with background information about the size and scale of libraries' operations, use of externally-created data, computer systems and formats, views on networking, etc., and
- the feasibility of a database with UNIMARC records from several sources.

Its findings were presented at a seminar held in Florence in June 1991.

The survey[1] of actual and potential use found a great number of different computers and operating systems; 12 national MARC formats + MAB (and no intention to abandon them), among them Italy using UNIMARC as national exchange format and Portugal as both input and exchange format; some agencies making UNIMARC versions of their records available, but only the Deutsche Bibliothek having conversion programs working in both directions; general agreement about the need for UNIMARC as the common exchange format in the EC, but many criticisms of it (often conflicting!); and wide support for a database network not restricted to EC member states. UNIMARC was not much used, although it was noted that co-operative projects, for example EROMM, some of them using CD-ROM, were taking shape and could well boost acceptance of the format. Booktrade organisations (with very few exceptions) made little response and revealed an alarming lack of awareness of UNIMARC and its potential uses.

---

1 A background analysis of MARC formats, carried out in 1994 by the UseMARCON project (see below), came to similar conclusions about the diversity of formats in use.

The feasibility study concluded that a small test database of UNIMARC records (including monographs, multi-volume publications and serials) from the national libraries of Belgium, France, Germany, Italy and Portugal could be feasible and could provide valuable data for analysis and comparison. The participants in the Florence seminar made several recommendations for:

- the improvement of the format,
- strengthening the hand of the IFLA UBCIM Office and the PUC;
- writing two-way national format-to-UNIMARC conversion programs as a matter of priority;
- using UNIMARC for all European co-operative bibliographic projects;
- using UNIMARC in retrospective cataloguing projects which involve converting old data into machine-readable form;
- establishing a network of databases; and
- a study of the feasibility of establishing a common database of authority files using UNIMARC/Authorities.

3.2 National Libraries Project on CD-ROM (CDBIB)

CDBIB had as its objectives "*to develop shared approaches to strategies, applications and formats for bibliographic data (especially national bibliographic data) on CD-ROM. This was designed to promote better and easier access by more users to European national bibliographies as well as promote economies in library cataloguing through an improved exchange of bibliographic records between European national agencies irrespective of different national MARC formats*". A major outcome of this project was a joint pilot disk ("The Explorers") containing records in a uniform UNIMARC format taken from the national bibliographies of Denmark (originally created in danMARC format), Italy and Portugal (two different implementations of UNIMARC) and the Netherlands (PICAMARC). It also produced research reports and specifications concerning

- MARC conversion routines between the UNIMARC format used in the pilot CD-ROM and the original MARC format
- European character sets
- Multi-lingual interfaces
- Links between CD-ROM and on-line hosts, and between CD-ROM and local library systems

The CDBIB project experimented with the development of conversion tools between MARC formats. At first the CCF Converter was considered for this purpose, but was found to be unsatisfactory, and the project developed its own prototype software for conversion. Testing and evaluation of the software showed that the approach - the use of modular, user-editable conversion tables as part of the conversion program - was generally sound, but that much more work would be needed to transform these results into a really satisfactory universal two-way conversion tool.

4. Ongoing projects & initiatives in Europe

4.1 UseMARCON (User Controlled Generic MARC Converter)

UseMARCON builds on the work of CDBIB and aims to complete by the end of 1996 or January 1997 at the latest. The conversion tool is designed to be used by senior cataloguers who have a good knowledge of MARC structure, possibly with some support from systems analysts. A graphical user interface for MS Windows or Unix Motif will make it possible for users to modify or create conversions by editing the conversion rules and tables.

The CDBIB project had already concluded that it would be impractical to attempt to create a table of every possible element which might be encountered in any format as the core of a one-step converter, so UNIMARC is used as the central switching format between any other pair of source and target formats (UNIMARC may, of course, itself be the source or target format). The reasons for the selection of UNIMARC as the core format were

- it offered a stable and maintained format
- to encourage its use

A by-product of the work has been the pin-pointing of elements in formats for which there are no UNIMARC equivalents, or the reverse. The basic data tables in UseMARCON cover UNIMARC, UKMARC, USMARC and InterMARC together with their corresponding character sets - UseMARCON is also designed to deal with the character sets which are designated for use with the various formats, and to make any necessary conversions.

4.2 CoBRA and CoBRA+ (Computerised Bibliographic Record Actions)

CoBRA was set up under the aegis of the Conference of European National Libraries (CENL) with funding under the European Commission's Libraries programme to promote discussion of core themes and technical issues regarding:

- improved European bibliographic services;
- user needs for bibliographic products;
- networked distribution and re-use of bibliographic records;
- data sharing between national bibliographic services, and
- longer term availability of electronic publications

In 1994 the European Commission funded five CoBRA initiatives, of which two are of prime interest to this workshop, the technical feasibility studies UNIMARC and AUTHOR. Another significant project is CHASE, also relevant to machine-readable bibliographic records, although not exclusively those in UNIMARC format.

The CoBRA concerted action has recently been extended as CoBRA+, whose key objectives build on and expand those of CoBRA, through Task Groups set up to address the following topics:

- metadata and bibliographic control and access with particular reference to electronic publications but not exclusively so;
- electronic publications and digital resources;
- exploitation of the results of CoBRA projects and actions, including their implementation and integration into library operations.

CoBRA UNIMARC

CoBRA UNIMARC is investigating the feasibility of UNIMARC to multinational databases. It is led by the Bavarian State Library working with a steering group taken from CERL, the 14-strong Consortium of European Research Libraries.

CERL. CERL was created following two conferences on retrospective cataloguing and conversion in 1990 and 1992, when a working party recommended the establishment of a common European database for the period 1450-1830, using UNIMARC as its format. CERL is assembling a wide spectrum of files from various sources, catalogued to differing standards over a very long period, some retroconverted and some modern original machine-readable records. Some files (from Lisbon and Zagreb) are created using UNIMARC; others (for example, those from ICCU, Rome) are derived from closely-related internal formats, but many are conversions from widely differing formats, including several variants of UKMARC from the British Library, MAB (Munich) and a local format using the STAIRS software package (Stockholm) which does not even conform to ISO 2709. After an international call for tenders, the Research Libraries Group (RLG) was selected as host for the CERL Hand Press Book (HPB) database. RLG uses its own version of USMARC, RLINMARC, so the records undergo a further process of conversion - and back again upon export from RLG. These are severe practical tests for UNIMARC. To date some half million records from the Bayerische Staatsbibliothek have been loaded; ca 49,000 18th century imprints from Stockholm, ca 3,300 from Zagreb and ca 40,000 from ICCU, Rome should be mounted by the end of the year. Several more from Paris, London, Madrid, Den Haag and Lisbon are in various stages of specification, analysis and testing.

The CERL files provide the CoBRA-UNIMARC study with a large body of data for analysis. The study aims to identify divergences in the use of the bibliographic description components of the UNIMARC record and to identify the problems arising when merging records from a number of sources. Particular objects of investigation were:

- the problems of differing interpretations of the options available in UNIMARC when merging records from multiple sources;
- the problems associated with holding, indexing and retrieving merged data from multi-lingual and multi-cultural sources; and
- the applicability of the minimum record content being prepared by the Permanent UNIMARC Committee across a merged database of records.

Conclusions about the second and third of these are largely conjectural (which does not mean they are invalid!), because there were long delays in preparing and

mounting files in the Consortium's HPB database, and the PUC had still not agreed a minimum record content by the time the project finished, although an interim version of it was seen.

The first part of the work was a statistical analysis of the files, for which special software was developed. This shows the fields and subfields which have been used, with their maximum, minimum and average lengths, and the presence of invalid fields and subfields; it also gives warning of general errors which would prevent correct analysis of the data (invalid record structures, etc.) and of inconsistencies indicating invalid UNIMARC data (missing mandatory fields, etc.). The software also produces an analysis of the character sets used in the files, making it easy to identify characters which are not part of the ISO standards prescribed in the UNIMARC manual. Character set issues indeed revealed themselves as one of the most difficult problems, and the software tool proved invaluable in helping to eradicate errors before files were sent to the USA for loading in the HPB database. This is powerful software which, although written as a DOS application for the specific purposes of this study, could be developed into a tool for use with any format.

The other part of the investigation was an "intellectual" (non-statistical) analysis of the content of sample records from the six files studied. This gives a description of the characteristic features of each file, in terms of both cataloguing rules and practice and also machine encoding or conversion, with examples of records, including a number of cross-file comparisons of records for the same items from different sources. Although the study highlighted several problem areas, for example the handling of multi-volume items, the outstanding finding has been the remarkable ability of UNIMARC to accommodate records created according to very different standards. Using UNIMARC is a balancing act: many of the alternatives built into the format are very useful, but their uncontrolled use can rapidly lead to needless inconsistencies and conflicts.

CoBRA - AUTHOR

The UNIMARC Study of 1990/91 and the Florence Seminar had recommended a study of the desirability and feasibility of setting up a common database of authority files in UNIMARC/Authorities format. CDBIB showed the feasibility of combining data from several sources with different formats, languages and cataloguing rules; although this was bibliographic data, it is no great step from this to authority data. CoBRA-UNIMARC has also commented on the wastefulness of records bearing authority data or links to national name authority files but whose information - painstakingly researched and authoritative - is not normally accessible to the library community and its users as a whole. National bibliographic agencies should be responsible for establishing the authoritative name forms for the persons and corporate bodies of their own political or linguistic areas, and the resulting data should be re- usable by other agencies and in public on-line databases.

Given the recognised importance or potential of authority data, project AUTHOR seeks to study the technical feasibility of giving access to authority files at the international level, converting authority data from the five national libraries which

are partners in the project to UNIMARC/Authorities format, and re-using the data in current cataloguing. To this end, AUTHOR will use the UseMARCON software to convert authority format records. It will also take into account results of the project EUROPAGATE in identifying a cost-effective and appropriate technological architecture.

Though the technical development and testing is still to be undertaken, valuable work has already been done on this project. AUTHOR can make a major contribution to progress in the field of authority control, with better appreciation and evaluation of authority file structures, not least UNIMARC/Authorities which has hardly been used in practice up to now, unlike its parent bibliographic format, and better utilisation of authority data.

CoBRA - CHASE : Character Set Standardisation

UNIMARC specifies the use of ISO 2022 and several ISO standard character sets, including ISO 646 (IRV) which is used as the basic control set (C0) and the default graphic set (G0) in all records. Switching between the latter and up to three further graphic sets is possible, if cumbersome; use of a wider range of graphic sets or characters which do not appear in the ISO standards can create considerable problems.

Project CHASE is designed to examine the feasibility of migration to Unicode for national bibliographic databases. The use of a unified character set encompassing all the characters needed in bibliographic records would seem at first sight to be the obvious way forward; there are, however, major problems in converting records from any earlier system of character sets to Unicode. These problems are by no means confined to UNIMARC: records in every MARC format would be equally affected, not only in their data content but also in their labels, directories and content designators which would all need to be changed.

5 Other Projects

5.1 Scanning and conversion

Two projects have looked at the general question of the conversion of printed bibliographic data into formatted, machine-readable records using scanning, OCR and SGML techniques, etc.

Project MORE worked on the conversion of printed library catalogues. It produced tested prototype workstation software configured to process any catalogue with a sufficiently homogeneous structure, allow for the display and editing of errors and uncertain characters, and convert the results to high quality UNIMARC formatted records. The final report of the project is available in French.

BiblioTECA has developed a toolbox for the analysis of the formal or informal structures which underlay not only catalogue records but also indexes, tables of contents and bibliographical references. The system can be run on a single PC or on a network of PCs, each handling a particular module. The products can be

MARC records in appropriate cases; UKMARC was used for some tests, but any other MARC format, including UNIMARC, could have been used.

5.2. SR / Z39.50

The other category of ongoing research projects which are confronted with record format issues are those implementing SR/Z39.50. In many cases, the source data are delivered from the servers (targets) in raw MARC format which then require conversion by the client to the local display format. Some, though not all, servers may offer clients a choice of export format. It is hoped that discussion will show whether results of the UNIMARC-oriented projects may be helpful also to other projects dealing in some way with the interactive exchange of bibliographic records.

Three projects which have been dealing with format issues in the networked environment are:

EUROPAGATE, a pilot gateway service through which users can access servers providing on-line access to catalogues; ONE (OPAC Network in Europe), which will link, amongst others, national libraries OPACs and includes modules for MARC conversion and character conversion; and SOCKER (SR Origin Communication KERnel), which is testing SR standards, focusing on the use of international networks, and query language translation. The testing will be done within three different environments (CD-ROM workstation, library system, and neutral access point).

Anthony G. Curwen
Aberystwyth, August 1996

FURTHER INFORMATION

Full Information on all Libraries Sector projects is available on

http://www.echo.lu/libraries/en/libraries.html

with links to project sites

Details about projects listed in this document:

UNIMARC-EC. Study to establish the feasibility of using UNIMARC amongst EC national libraries, bibliographic utilities and the booktrade based upon their present computer facilities: final report in English and German by Die Deutsche Bibliothek; with appendices by Anthony G. Curwen, Consultant, UK and Trudi C. Noordermeer, The Royal Library, NL (CEC, December 1992).

CDBIB - National Libraries Project on CD-ROM: Jointly funded by the Consortium of National Libraries and DG XIII/E under its IMPACT programme. Web site: http://www.konbib.nl/kb/sbo/proj/cdbib/

UseMARCON - User Controlled Generic MARC Converter
Contact: Drs Trudi Noordermeer, Koninklijke Bibliotheek
PO Box 90407, NL-2509 LK Den Haag
Tel: +31 703140 597 Fax: +31 703140 424 e-mail: trudi.noordermeer@python.konbib.nl

CERL - Consortium of European Research Libraries
Contact: Mr Bob Henderson, Project Manager, The British Library
Great Russell Street, UK - London WCIB 3DG
Tel: +44 171 412 7073 Fax: +44 171 412 7563 e-mail: bob.henderson@bl.uk
Web site: http://portico.bl.uk/cerl/

CoBRA-UNIMARC: Feasibility of the application of UNIMARC to multinational databases
Contact: Dr Claudia Fabian, Bayerische Staatsbibliothek
Ludwigstrasse 16, D-80328 München
Tel: +49 89 28638 323 Fax: +49 89 28638 293
e-mail: 101473.101@compuserve.com

CoBRA-AUTHOR: Feasibility study into the transnational application of national name authority files
Contact: Francoise Bourdon or Sonia Zillhardt
Bibliothèque Nationale de France, Direction du développement scientifique et des réseaux, 2, rue Vivienne, F-75002 Paris Cedex 02
Tel: +33 1 47 03 86 46 (Bourdon); +33 1 47 03 77 08 (Zillhardt)
Fax +33 1 47 03 81 50 e-mail: francoise.bourdon@bnf.fr sonia.zillhardt@bnf.fr

CoBRA-CHASE: Character set standardisation - migration strategies to Unicode for national bibliographic databases
Contact: Mr Anthony Brickell,The British Library, 2 Sheraton Street,
UK- LONDON WIV 4BH. e-mail: anthony.brickell@bl.uk

CoBRA+ Computerised Bibliographic Record Actions Plus Preservation and Service Developments for Electronic Publications
Contact: Mr Howard Shoemark,The British Library
Boston Spa, Wetherby, UK - LS23 7BQ
Tel: +44 1937 546596 Fax: +44 1937 546586 e-mail: howard.shoemark@bl.uk

MORE - MARC Optical REcognition
Contact: Mme C. Lupovici, Chef de Produits Bibliothèques, Jouve SI
18, rue Saint-Denis, BP 414-01, F-75025 Paris Cedex 01
Tel: +33 1 44 76 86 17 Fax: +33 1 44 76 86 10 e-mail: clupovici@jouve.fr

BiblioTECA
Contact: Mr Jaime Sarabia, Universidad Complutense de Madrid, Filosofia B, Laboratorio de Inteligencia Artificial, Ciudad Universitaria, E-28040 Madrid
Tel: +34 1 394-60-54 Fax: +34 1 394-60-53 e-mail: sarabia@eucmax.sim.ucm.es
Web site: http://www.csic.es/cbic/teca.htm

EUROPAGATE-
Contact:Ms Annette Kelly, Library Council of Ireland,
53-54 Upper Mount Street, Dublin 2
Tel: +353 1 67 61 167 or +353 1 67 61 963 Fax: +353 1 67 66 721
e-mail: annkelly@tco.ie

ONE - OPAC Network in Europe
Contact: Fru Liv Holm, BRODD, Oslo College, Pilestredet 52
N-0167 Oslo
Tel: +47 22 45 26 00 Fax: +47 22 45 26 05 e-mail: liv.a.holm@brodd.hioblo.no
Web site: http://www.bibsys.no/one.ta.html

SOCKER - SR Origin Communication KERnel
Contact: Mr Arne Sorensen, UNI-C, Oluf Palmes Alle 38, DK-8200 Arhus
Tel: +45 86 78 44 44 Fax: +45 86 78 44 55 e-mail: recas@ums2.uni-c.dk
Web site: http://mediator.uni-c.dk/socker

AGENDA

UNIMARC WORKSHOP

Luxembourg, Room Glesener A, Wagner Building

Friday 13.09.1996

Session I: Taking Stock

9.30 Welcome and introduction. Purpose and objectives of the meeting (A. Iljon)

9.50 Barriers to international bibliographic record exchange and progress on removing format incompatibilities: Overview and context of Commission actions (P. Manson)

10.10 Presentation of available solutions through demonstration of project results and discussion of their impact and take-up: Projects UseMARCON, UNIMARC, AUTHOR

11.30 Tour de table of other projects (ONE, CHASE, KSYSERROR, SOCKER, BIBLINK) followed by discussion

13.00 Lunch

Session II: Looking Ahead

14.00 Important developments: The initiatives of IFLA and OCLC

15.00 The prevention of the development of format barriers in the future exchange of extended bibliographic information in the electronic document environment

15.45 Rapporteur's summary of main issues. Discussion and suggestions of actions and recommendations needed to sustain and continue the process of overcoming format barriers

16.30 Summary and Conclusions (A. Iljon)

17.00 End of Workshop

List of Participants

UNIMARC WORKSHOP

13-09-96

BELGIUM

Mrs. Paula GOOSSENS
Bibliothèque Royale Albert Ier
(Koninklijke Bibliotheek)
Boulevard de l'Empereur 4
B - 1000 BRUXELLES
Tel: 25195648
Fax: 25195646

DENMARK

Mr. Erik BERTELSEN
Head of Research & Development
UNI.C
Olof Palmes Allé 38
DK-8200 AARHUS N
Tel: 893766
Fax: 89376677
E-Mail: erik.bertelsen@uni-c.dk

Mrs. Elise HERMANN
Statens Bibliotekstjeneste
Nyhavn 31E
DK-1051 KØBENHAVN
Tel: 33934633
Fax: 33936093

Mrs. Susanne THORBORG
Dansk BiblioteksCenter a.s.
Tempovej 7-11
DK - 2750 BALLERUP
Tel: + 45/44867832
Fax: + 45/45867863
E-Mail: st@dan.bib.dk

FINLAND

Ms. Eeva MURTOMAA
Helsinki University Library
Teollisuuskatu 23
P.O.Box 26
FIN - 00014 HELSINKI
Tel: 070844318
Fax: 070844341
E-Mail: eeva.murtomaa@helsinki.fi

FRANCE

Mrs Françoise BOURDON
Responsable de la Mission pour l'Organisation
Direction du Développement Scientifique et des Réseaux
Bibliothèque Nationale de France
Rue Vivienne 2
F - 75081 PARIS CEDEX 02
Tel: + 33/147037708
Fax: + 33/147038150

Mrs. Catherine LUPOVICI
Ingénieur Consultant/Chef de Produits Bibliothèques
Jouve Systèmes d'Information
Rue Saint-Denis 18
B.P.2734
F-75027 PARIS CEDEX 01
Tel: 144768617
Fax: 144768610

Ms. Florence ROBERT
Agence Bibliographique de l'Enseignement Supérieur
Rue Guillaume Dupuytren 25
Le Florence-Parc Euromédecine
B.P.4367
F - 34196 MONTPELLIER CEDEX 5
Tel: + 33/67548410
Fax: + 33/67548414
E-Mail: robert.florence@abes.fr

Mrs. Sonia ZILLHARDT
Responsable de la Coopération Européenne
Bibliothèque Nationale de France
Rue Vivienne 2
F-75081 PARIS CEDEX 02
Tel: 147037708
Fax: 147038150
E-Mail: sonia.zillhardt@bnf.fr

GERMANY

Mrs. Claudia FABIAN
Bayerische Staatsbibliothek
Ludwigstrasse 16
Postfach 340150
D - 80539 MÜNCHEN
Tel: 8928638331
Fax: 8928638293

Dr. Volker HENZE
Standard officer
Die Deutsche Bibliothek
Zeppelinallee 4-8
D - 60325 FRANKFURT/Main
Tel: 697566729
Fax: 697566709

Mrs. M.F. PLASSARD
Die Deutsche Bibliothek
Zeppelinallee 4-8
D-60325 FRANKFURT/Main
Tel: 697566201
Fax: 697566476

GREECE

Mrs. Zinovia PAPAGEORGIOU
National Library of Greece
Panapistimiou Street 32
GR - 10679 ATHENS
Tel: + 30/13608495
Fax: + 30/13611552

IRELAND

Mr. Brian McKENNA
National Library of Ireland
Kildare Street
IRL - DUBLIN 2
Tel: 166118811
Fax: 16766690

ITALY

Dr. Isa DE PINEDO
Head of dept.
Istituto Centrale
Catalogo Unico delle Biblioteche Italiane
Informazioni Bibliografiche
Via del Castro Pretorio 105
I - 00185 ROMA
Tel: 64989482
Fax: 64959302
E-Mail: de Pinedo @ ITCASPUR.CASPUR.IT

LUXEMBOURG

Mr. Emile THOMA
Conservateur
Bibliothèque nationale
37, boulevard F.-D. Roosevelt
L - 2450 Luxembourg
Tel: 229755244
Fax: 475672
E-Mail: thoma@bi.etat.lu

NETHERLANDS

Mrs. Trudi NOORDERMEER
Research Librarian
Koninklijke Bibliotheek Nederland
National Library of the Netherlands
Prins Willem-Alexanderhof 5
P.O.Box 90407
NL-2509 LK DEN HAAG
Tel: 703140597
Fax: 703140424
E-Mail: trudi.noordermeer @ konbib.nl

Mr. M. van MUYEN
Director
Centrum voor Bibliotheekautomatisering
PICA
Schipholweg 99
P.O.Box 876
NL-2300 AW LEIDEN
Tel: 71257168
Fax: 71223119

NORWAY

Ms Annema H. LANGBALLE
Senior Academic Librarian
National Library of Norway
Bygdøy Allé 21
Solli
P.O.Box 2674
N - 0203 OSLO
Tel: 22553370
Fax: 22553895

PORTUGAL

Dra. Fernanda M. CAMPOS
Vice-Presidente
Instituto da Biblioteca Nacional e do Livro
Campo Grande 83
P-1700 LISBOA
Tel: 17950130
Fax: 17933607

SWEDEN

Mr. Christer LARSSON
Bibliographic Projects Manager
Libris Dept.
The Royal Library Stockholm
Kungl.Biblioteket
Box 5039
S - 10241 STOCKHOLM
Tel: 84634262
Fax: 84634265
E-Mail: christer.larsson@libris.kb.se

UNITED KINGDOM

Mr. Antony G. CURWEN
Bodnant Primrose Hill
Aberystwyth
Llanbadarn Fawr
UK- DYFED SY23 3SE
Tel: 1970611861
Fax: 1970611861

Mr. Michael DAY
University of Bath
UK- Office for Library Networking
Claverton Down
UK - BATH BA2 7AY
Tel: 1225826580
Fax: 1225826838
E-Mail: UKOLN@Bath.ac.uk

Mr. Bob HENDERSON
Project Manager
The British Library
Humanities & Social Sciences
Great Russell Street
UK - LONDON WC1B 3DG
Tel: 1714127581
Fax: 1714127762

Mr. Brian P. HOLT
The British Library
Boston Spa
Wetherby
UK - WEST YORKSHIRE LS23 7BQ
Tel: + 44/193754696
Fax: + 44/193754696

Mrs. Janet MITCHELL
Director Europe
OCLC International
Hagley Road 51-53
Edgbaston
7th Floor Tricorn House
UK- BIRMINGHAM B16 8TP
Tel: 1214564656
Fax: 1214564680

Mr. Peter SMITH
Deputy Director
London & South Eastern Library Region
Wapping Lane 70
Fourth Floor-Gun Court
UK - LONDON E1 9RL
Tel: 1717022020
Fax: 1717022019

Mr. Neil WILSON
National Bibliographic Service
The British Library
Boston Spa
Wetherby
UK - WEST YORKSHIRE LS23 7BQ
Tel: + 44/1937546585
Fax: + 44/1937546586
E-Mail: neil.wilson@bl.uk

USA

Mrs. Marda JOHNSON
Online Computer Library Center
Frantz Road 6565
USA - DUBLIN - OHIO 43017
Tel: 6147646000
Fax: 6147646096

ANNEX II

TECHNICAL EXPERIENCES OF UNIMARC AND CONVERSIONS:

Papers presented:

1. UseMARCON
2. CoBRA/UNIMARC
3. CoBRA/AUTHOR
4. OCLC UNIMARC Development: a status report

USEMARCON
a generic convertor for MARC formats

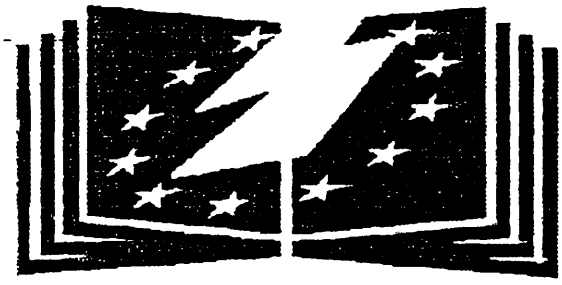
UNIMARC Workshop
Luxembourg
13 September 1996

Koninklijke Bibliotheek
Library Research Department

trudi.python@konbib.nl

USER CONTROLLED GENERIC MARC CONVERTOR

USEMARCON



Factsheet - July 1996

WHY IS THE USEMARCON PROJECT NEEDED?

Different national MARC (MAchine Readable Catalogue) standards are seen as barriers to wider exchange of bibliographic records in Europe and beyond. Throughout the world nearly 50 different MARC formats are currently in use, with 10 employed in the national libraries of European Community countries. Such variation is a fundamental problem for libraries wishing to obtain or supply data internationally and often results in the re-cataloguing of material for which records are readily available in formats other than the library's own. Lack of language expertise or knowledge of the context of publication can lead to records of a significantly inferior quality to those which might otherwise have been obtained from an agency in the country of publication.

WHAT ARE THE OBJECTIVES OF THE USEMARCON PROJECT?

Development of a generic MARC record conversion system to enable libraries to easily convert records between the various national MARC formats.

To give libraries the ability to obtain records from a far wider range of potential sources than those currently available to them.

Stimulate an increase in the international exchange of bibliographic records.

WHO IS FUNDING USEMARCON?

The USEMARCON Project is funded by the consortium partners and the EU's Telematics Applications Programme (DGXIII-E).

HOW WILL THE TECHNICAL OBJECTIVES OF USEMARCON BE REALISED?

The USEMARCON software application is a highly versatile rules based conversion program capable of running in either the MS Windows/MS Windows 95' or Unix (Solaris)/Motif operating environments. The modular construction of the program will allow varying levels of conversion to be performed ranging from simple character set translation to complete conversions between different MARC formats (e.g. UKMARC, UNIMARC, USMARC etc.). In order to allow the program to be as flexible as possible, users will be provided with the ability to customise or create conversions to match their local requirements by the editing of ASCII rules files and conversion tables.

The program itself will be developed using an object-oriented methodology and written in the C++ programming language using the XVT cross platform development toolkit.

WHO IS INVOLVED IN THE USEMARCON PROJECT?

The partners of the USEMARCON Project consortium are drawn from a variety of library and information technology backgrounds and comprise the following:

Co-ordinating Partner
Koninklijke Bibliotheek, Holland

Full Partners
Instituto da Biblioteca Nacional e do Livro, Portugal
The British Library, UK

Associate Partner
Die Deutsche Bibliothek, Germany

Software Developer
Jouve, Systèmes d'Information, France

WHAT PROGRESS HAS BEEN MADE SO FAR?

Prior to development of the conversion program a technical and commercial feasibility study was undertaken in 1994. Following this study an inventory of potential conversion problems was made and several available packages with MARC conversion capability were evaluated. Successful completion of this study led to initiation of the second stage of the project by the consortium in March 1995

A pre-design phase resulted in the delivery by Jouve of a global functional design defining data structures and the conversion instruction set in September 1995. The first alpha version of the software was delivered in February 1996. Testing of the software by consortium partners is currently underway using the InterMARC, UKMARC and UNIMARC bibliographic and authority formats. It is planned to implement USMARC conversions before the end of the project in October 1996.

The consortium plans to conduct market research from July to September 1996 concerning the possibility of commercial exploitation of the results of USEMARCON with a view to developing a range of products.

HOW CAN I FIND OUT MORE?

As the work of USEMARCON proceeds further information will be provided through special mailings and press releases.

If you wish to receive details of progress please contact:

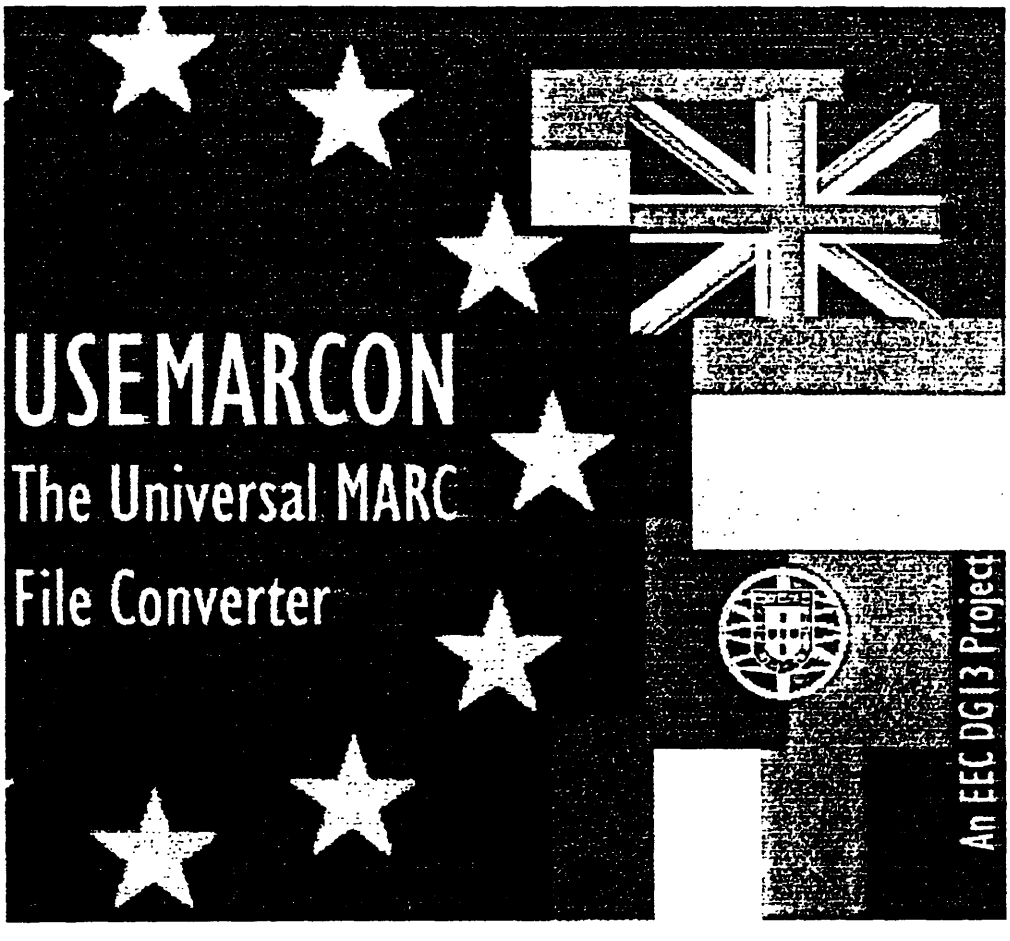
Trudi Noordermeer
Koninklijke Bibliotheek
Library Research Department
PO Box 90407
2509 The Hague,
The Netherlands

Phone: +31 70 3140597
Fax: +31 70 3140424
Email: trudi.@python.konbib.nl

USER CONTROLLED
GENERIC MARC CONVERTER

USEMARCON

Technical Description - July 1996



Introduction

The USEMARCON system is designed to allow users to convert bibliographic records from any ISO 2709 compatible MARC (MAchine Readable Catalogue) format (e.g UKMARC) to any other (e.g. InterMARC) using UNIMARC as a central switching format. USEMARCON does not require programming experience and is designed to be used by senior cataloguers with a good knowledge of MARC structure. The USEMARCON prototype includes basic tables for UNIMARC, UKMARC, USMARC and InterMARC together with their corresponding character sets. In order to allow the program to be as flexible as possible, users are provided with the ability to customise or create conversions to match their local requirements by the editing of ASCII rules files and conversion tables.

The USEMARCON Graphical User Interface (GUI) is designed for use with both MS Windows (3.1x and '95) and Unix Motif environments, in the latter case running under the Sunsoft Solaris operating system. In the Windows environment USEMARCON uses a Multiple Display Interface.

Progress Of The Project

Details of progress will be provided through special mailings and press releases. If you wish to receive further information please contact:

Trudi Noordermeer
Koninklijke Bibliotheek
Library Research Department
PO Box 90407
2509 The Hague
The Netherlands

Phone: +31 70 3140597
Fax: +31 70 3140424
Email: trudi.@python.konbib.nl

Basic Concepts

Bibliographic data conversion means the processing of:

- MARC bibliographic data, including related blocking and fill characters.
- Conversion rules for specifying the conversion of each data element of the input format into each data element of the output format.
- Coded data tables corresponding to sets of codes used by the format, eg. language of publication, country of publication.
- Character set tables specifying the translation of individual character codes between input and output formats.
- Format checking tables specifying valid elements of input and output formats.

MARC Formats[1]

The MARC format family was created through a Library of Congress project, initiated in 1964, to prepare bibliographic information for automated processing. Different national MARC formats were created in response to specific national needs according to their local cataloguing rules and operating environments. The principles of MARC formats are to structure the bibliographic data into fields and sub-fields as in the following UNIMARC example :

Field tag	Indicators	Sub-field delimiter	Title proper	Sub-field delimiter	Subtitle
200	**1b**	**$a**	**UNIMARC**	**$e**	**Cataloguing Manual**

Indicators are numeric values used for specific processing of the field, where there is no specific value they are replaced by a blank. Sub-field delimiters are generally introduced by $ followed by a letter or a number qualifying the data element in the field, ($a, $b or $1, $2). Some fields contain a single sub-field with the data coded in a fixed position and length inside. The format specifies whether each field and sub-field are mandatory, optional or conditional upon another data element. USEMARCON enables MARC formats to be described in tables which can be used to control the input and the output of data. The tables have the structure given in the following UNIMARC example showing the list of fields and sub-fields with occurrence qualifiers together with possible values of indicators.

```
100_  |I1=_   |I2=_   |$a_
200_  |I1=01  |I2=_   |$a+ |$b* |$c*            // Title
700*  |I1=_   |I2=012 |$a_ |$b? |$c? |$d? |$f? // Author
```

_ = *mandatory not repeatable* ? = *optional non-repeatable*
\+ = *mandatory & repeatable* * = *optional & repeatable*

The USEMARCON data input file must comply with the ISO 2709 bibliographic data standard to be loaded and processed by the system. This format is also commonly known as 'MARC Communications Format'. Output data produced will also be fully compliant with ISO 2709. Users are provided additionally with the ability to specify particular details of the data structure used for input and output, including: blocking factor, segmentation, minimum size of usable data blocks and specification of padding character.

Character Set Tables

Bibliographic information is coded using extended character sets to cover a large range of Latin and non-Latin scripts. As different MARC formats use different character sets, the reformatting process offered by USEMARCON includes character set conversion. This is handled by the use of tables mapping input to output character sets which can be edited to meet local requirements. The structure of a character set conversion table is shown in the following example converting from ISO 5426 to ASCII :

```
ISO5426
ascii
#include "basic.trf"
0xA3   |    0x9C  // Pound character - simple conversion
0xCA   |    0xF8  // Degree - if 0xCA appears before an A see below
0xCA A |    0x8F  // A with circle above
0xBF   |          // The inverted question mark is not converted
```

Other 'lookup tables' eg. country of publication code, language code etc can be created for use in combination with a main rules file. These have the same structure as the character set conversion file and can be edited with any text processor.

[1] Campos, Fernanda M. ; Lopes, M. Inês ; Galvão, Rosa M. - Marc formats and their use : an overview. In : *Program*, vol. 29, n° 4, October 1995, pp. 445-459

Rules For Format Conversion

For the purpose of writing rules for data conversion, the input and output data parts are named with a specific Content Designator (CD) based on the corresponding MARC format. The conversion rules describe how each input CD or each set of input CDs are converted into each output CD. The rule syntax includes a set of basic operators and allows six types of instruction to be used : conditional, boolean, loop, memory, conversion and translation. All the rules for a specific MARC conversion are gathered in a single file, which can be created or edited either with the USEMARCON rule editing tool or with any text processing program. A simple rule file structure is followed. The structure for title field conversion from UNIMARC to UKMARC is shown in the following example:

```
200I1       | 245I1       If (S=0) Then S
            | 245I1       If (S=1 And (Exists(700) Or Exists(710) Or Exists (720))) Then 1
            | 245I1       If (S=1 And Not(Exists(700) Or Exists(710) Or Exists (720))) Then 3
200$a       | 245I2       |  Sto(0); Bfirst('\88'); Sto(1); Mem(0); Bfirst('\89'); S-Mem(1)
            | 245$a       If (n=1) Then Delete('\88'); Delete('\89');
            | 245$i       If (n>1) Then S
200$b       | 245$z
200$c       | 245$j
200$d       | 245$k
200$f       | 245$e       // First $e created
200$g       | 245$e       // Subsequent $es created
200$h       | 245$l(no)   Sto(0); Next($i,$h);
                          If (S='') Then Mem(0) Else Mem(0)+', '+S
                          // Searches next $i until next $h
                          // If found merges it with $h+', '
                          // else only $h in $l
200$i       | 245$l(no)   Sto(0); Last($h,$i);
                          If (S='') Then Mem(0)
                          // Searches last $h until last $i
                          // If not found then create a new $l containing only $i
                          // If found there is nothing to do, the precedant rule has
                          // already create the subfield

200$h       248I1         | ns
            248I2         | 0
200$h       248$g         // first h, I or hi combination has 248 I1=1, second has I1=2, etc.
200$i       248$h         // The routine for separating out combinations is as above
```

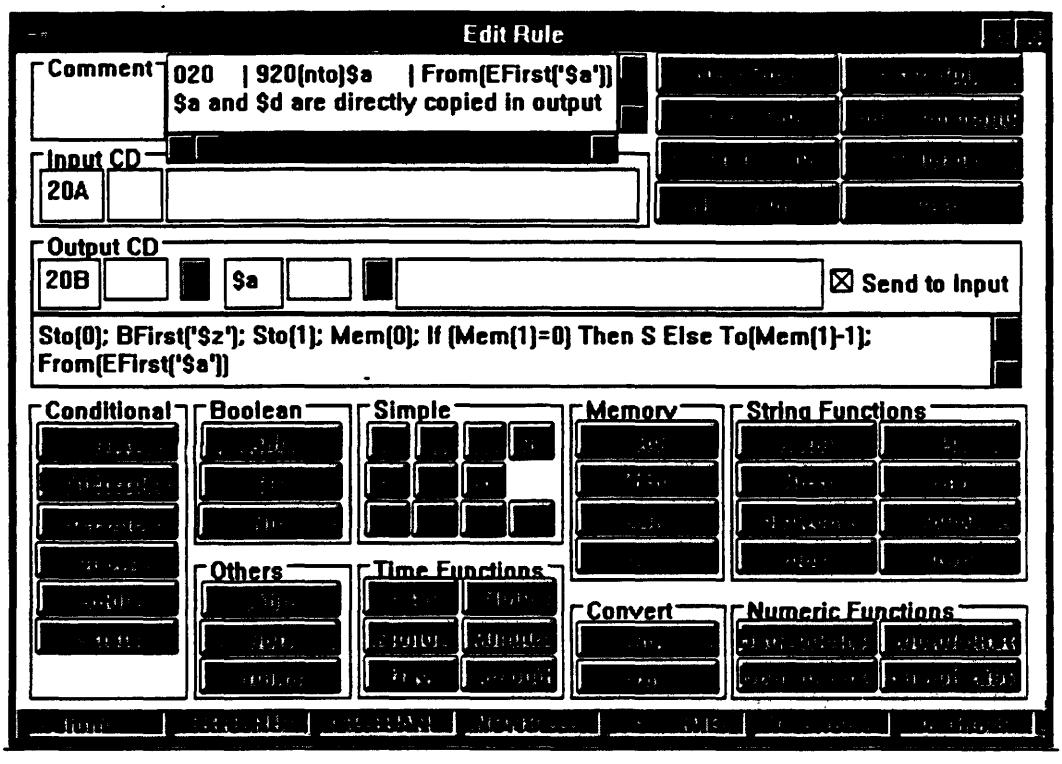
System Functionality

The USEMARCON system allows rules used by the conversion process to be created, edited and tested. Conversions can be run in either a step-by-step 'interactive' or batch mode. Interactive mode is particularly useful when testing rules or tables to check the accuracy of conversions before undertaking a large batch mode conversion. Editing can be done either using the USEMARCON editor or a regular text editor.

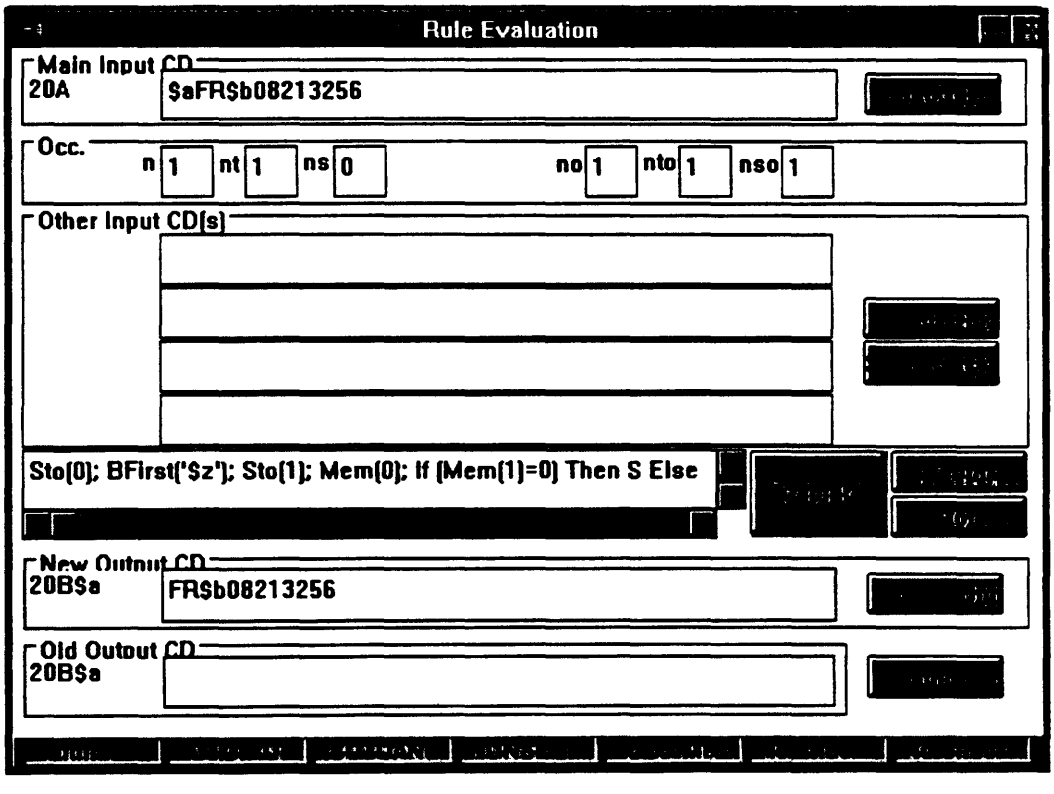
Editing Mode

The USEMARCON rules editor provides an interface allowing the creation and editing of rules linking input and corresponding output CDs. All the commands which can be used in rules can be displayed and automatically activated by the use of buttons.

On-line help provides explanations and examples of the syntax for writing the rules.



Users may also evaluate a rule against loaded or keyed CDs.



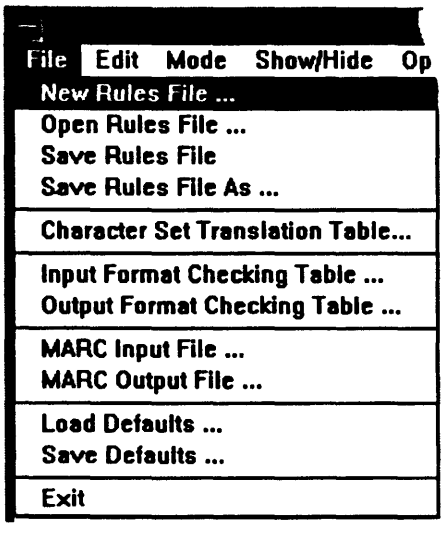
According to the number of CDs used in the rule being evaluated, one or more input CDs may be used. Occurrences where rule evaluation has identified potential problems can easily be amended via the program interface.

Rules are processed according to the following table :

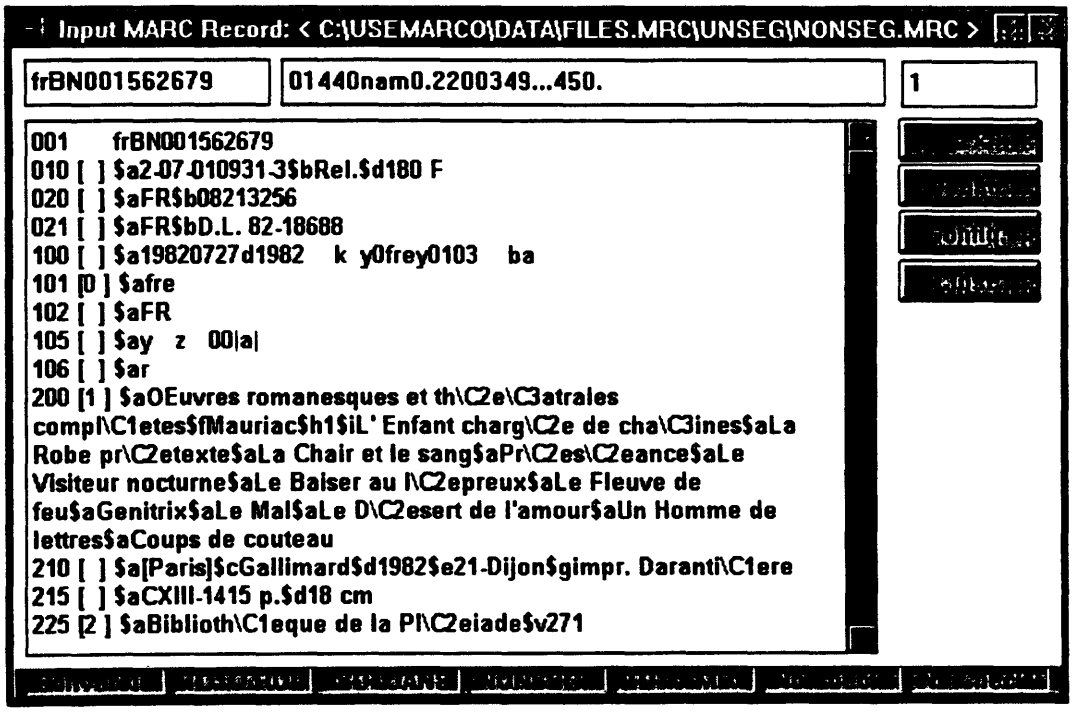
CD In		CD Out				B	
Tag	Sub-field	Tag	Occ	Sub-field	Occ	Rule	Comment
NR	NR	NR		NR			The CDIn goes in CDOut
NR	NR	NR		R			Idem, only one occurrence of CDOut will exist
NR	NR	R		NR			Idem, only one occurrence of CDOut will exist
NR	NR	R		R			Idem, only one occurrence of CDOut will exist
NR	R	NR		NR		+ -	Each sub-field of CDIn will go in the same sub-field of CDOut (they must be merged : + at beginning of rule signifies Destination+...).
NR	R	NR		R	*n*		Each occurrence of sub-field in input will create a new occurrence of sub-field in output.
NR	R	R	*n*	NR			Each occurrence of sub-field in input will create a new occurrence of field in output.
NR	R	R		R	*n*		Each occurrence of sub-field in input will create a new occurrence of sub-field in output
NR	R	R	*n*	R			Each occurrence of sub-field in input will create a new occurrence of field in output.
R	NR	NR		NR		+	Each occurrence of field in input will be merged in the same sub-field (if + is omitted at the beginning of the rule, an error of format can occur in output).
R	NR	NR		R	*n*		Each occurrence of field in input will create a new subfield in the same field in output
R	NR	R	*n*	NR			Each occurrence of field in input will create a new occurrence of the field in output
R	NR	R		R	*n*		Each occurrence of field in input will create a new subfield in the same field in output
R	NR	R	*n*	R			Each occurrence of field in input will create a new occurrence of the field in output
R	R	NR		NR		+	Each occurrence of field and/or sub-field will be merged in the same sub-field in output (if + is omitted at the beginning of the rule, an error of format can occur).
R	R	NR		R	*n*		Each occurrence of field and/or sub-field in input will create a new occurrence of sub-field in output
R	R	R	*n*	NR			Each occurrence of field and/or sub-field in input will create a new occurrence of field in output
R	R	R		R	*n*		Each occurrence of field and/or sub-field in input will create a new occurrence of sub-field in output
R	R	R	*n*	R			Each occurrence of field and/or sub-field in input will create a new occurrence of field in output
R	R	R	*nt*	R	*ns*		Each occurrence of field (numbered *nt*) will create a new occurrence of field. Each occurrence of sub-field (numbered *ns*) will create a new occurrence of sub-field within current field.

Conversion Mode

The conversion function allows the specification of an input data file whose converted records will be stored in an output data file specified by the user. The conversion process uses the rule file defined for each conversion and, if appropriate, character set conversion tables and MARC format checking tables for input and output files. Before processing data users select the rules and tables files necessary via the main file menu.



The conversion can then be processed in either interactive or batch mode. Interactive mode allows display of either or both the input and output records. The USEMARCON software enables additionally output records to be edited manually to overcome specific problems which cannot be handled through the general conversion rules. Edited records can be saved in an updated MARC output file.



Diacritics and other extended characters are always displayed in hexadecimal code in order to allow users to check the accuracy of character translation

In batch mode or interactive modes details of processing problems are stored in a report file and classified by the following error types: input format, MARC checking in input, character translat ion, coded data translation, conversion, MARC checking in output and building of the output MARC file.

/Ends

PROBLEM DOMAIN

Conversion of MARC formats : one of the basic problems for the exchange of bibliographic data

Worldwide more than 50 MARC formats in use

Objective USEMARCON
To develop a generic convertor for MARC formats (real ISO 2709, which excludes e.g. Pica and MAB)

MARC formats used by national libraries in the CEC :

Country :	MARC formats :
Belgium	InterMARC and UNIMARC
Denmark	danMARC
France	InterMARC UNIMARC
Germany	MAB1 / UNIMARC
Greece	UNIMARC
Ireland	UKMARC
Italy	UNIMARC
Luxembourg	SIBILMARC
Portugal	UNIMARC
Spain	IBERMARC / UNIMARC (?)
Netherlands	PicaPlus USMARC
U.K.	UKMARC

BLCMPMARC
NORMARC
SLSMARC
CANMARC
ADABAS/WINMARC
AUSMARC
CHINAMARC/CNMARC
CSMARC
JAPMARC
LCMARC
LIBRISMARC
MALMARC
PHILMARC
UBVUMARC Univers. Amsterdam
WILSONMARC
BNBMARC
SweMARC
CatMARC
PicaPlus
AnnaMARC
Mekof

PLANNING

Idea : December 1992

Phases:
1. feasibility study
2. development Alpha version
3. extended testing/development and documentation

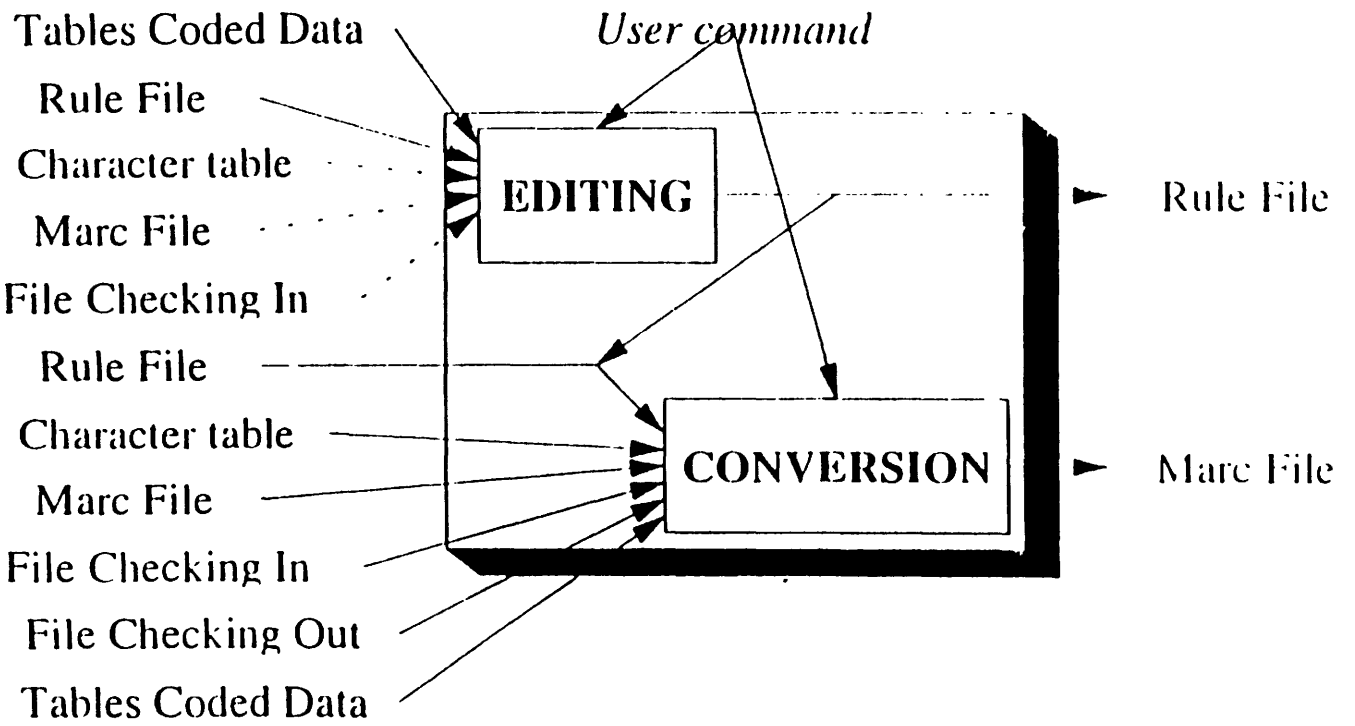
Start phase 1 : February 1994
End phase 1 : 6 October 1994

Intermediate period

Start phase 2 : March 1995
Delivery alpha version : May 1996

Start phase 3 : May 1996
Final delivery convertor October 1996

Global functional architecture



EXAMPLE OF USE:

Convert DANMARC records to SWEMARC records

1 Conversion table DANMARC -> UNIMARC
2. Conversion table UNIMARC -> SWEMARC
   - UNIMARC does not have an equivalent for all subfields in other formats;
   - what to do when fields have to be split?
3. Adaptation of rules files
4. Description of DANMARC for check input records
5. Description of SWEMARC for check output records

DELIVERABLES

1. Conversion software

2. Conversion tables :
   UKMARC -> UNI -> UKMARC
   USMARC -> UNI -> USMARC
   InterMARC ->UNI-> InterMARC
   (UNIMARC is used as the central format)

3. Format descriptions of USMARC, UNIMARC, InterMARC & UK MARC for format checking

4. Set of conversion rules

5. User Documentation

6. Technical Documentation

DOCUMENTATION

1. User manual

2. Technical manual

3. Format descriptions (from the 'owners' of the formats

4. Conversion tables (from libraries / library automation companies etc. interested in conversions)

5. Sets of format conversion rules

IMPACTS AND BENEFITS

Conversion of one MARC format to another MARC format is an expensive operation. A generic convertor which is relatively easy to use and to adapt is useful to overcome barriers in the exchange of bibliographic records

Very ambitious project, because of practicle maintenance problems
Earlier attempts, e.g.
National library of Canada
National Library of Australia and
CCF convertor

PROBLEMS SOLVED

1. USEMARCON is a useful tool for the conversion of MARC formats and it can be used rather easily and it can be adapted for other conversions rather easily

PROBLEMS REMAINED (logical and practical)

1. It is not possible to convert all possible fields because sometimes there are no corresponding fields

2. Keep format descriptions up-to-date in the organisations who own the format : very difficult
   - formats change all the time (e.g. adaptation for description of electronic publications)

- format descriptions are not available (Pica)
- in a langauge which many people can't read (DanMARC)
- obsolete descriptions:InterMARC

Solution could be :
national libraries (and CENL) is responsible for keeping the national format description up-to-date (in national language AND in English)

3. Putting USEMARCON in the market:
   - marketing
   - distribution
   - help desk
   - training/demonstrations

   No budget available for maintenance and exploitation.

4. **Which package is distributed?**
   - **Executables of the convertor**
   - **Source code**
   - **Conversion tables**

   How to organise updates?

   USEMARCON software must be maintained otherwise it will not be useful anymore on the long term

5. **'Price' (can be for free) of USEMARCON**
   - **for the project partners**
   - **for the associate partner**
   - **for other EC projects**
   - **for libraries**
   - **for commercial companies**

OTHER RELEVANT DEVELOPMENTS:

1. **Much interest in online conversion: USEMARCON could be adapted for this, but it is expensive and there is no budget for it**

2. **Format integration of USMARC, UKMARC and CANMARC**
   Majority of bibliographic MARC records will be based in one of these three or in a format which looks like UKMARC or USMARC

UNIMARC REPORT

UNIMARC WORKSHOP - LUXEMBOURG, FRIDAY 13TH SEPTEMBER 1996

Dr. Claudia Fabian - Bayerische Staatsbibliothek

When the UNIMARC project (more correctly the study concerning the "Feasibility of the application of UNIMARC to multinational databases") was proposed to the European Commission (EC) for funding, the Bavarian State Library (BSB) as co-ordinator and all participants were determined and motivated by an extremely practical and highly important and useful background. In fact the database planned by the Consortium of European Research Libraries CERL for early printed books (so-called Hand Press Book, HPB-database) was to be fundamentally UNIMARC-based to accept and foster the role of UNIMARC as a commonly agreed European exchange format.

Since January 1995, the BSB is the contractor for this EC project, which has been defined by CERL as desirable for and conducive to European co-operation. To meet Commission requirements, this project was assigned to a European library, not to the Consortium. The COBRA-UNIMARC Project is the first EC project of the BSB, and in fact (although not juridically) the first EC project of CERL as well. CERL decisively supports the propagation and active application of UNIMARC by obliging all libraries to input their data using this particular format. It thus helps to build a stronger base of support for UNIMARC - and unless UNIMARC is increasingly widely adopted, it may fail to become a truly internationally used format. Ideally the data in the HPB-database should be kept in UNIMARC, but that could not be realised in the first phase.

The aim of the project was to identify and test the problems arising from differing interpretations of the options available in UNIMARC for merging records from multiple sources, the problems associated with holding, indexing and retrieving merged data from multilingual and multicultural sources and to study the applicability of a minimum record content across a merged database of records. These issues were bound to become vital for the building up and functioning of the HPB-database, unless they were detected and prevented or maybe eliminated before. CERL is building up a coherent database of European imprints until about 1830, as a source of records for both cataloguers from everywhere and international research purposes. This precisely defined project permits a practical test of the chances and limits of European library co-operation, as it is not enough to define principles and pronounce statements of intent, which are quickly gauged by reality when compared with concrete results. This project demonstrates the importance of European and international library-oriented standards, and the need to assess the deficits inherent in their definition or evident in their realization. It is not that CERL is revolutionary - it is the first real chance of European libraries to discuss and concretely experiment issues they must all face in the future. Here is an organisation which must prove or test the feasibility of European co-operation, but also to become aware of our special European needs and to compare our advantages and difficulties with those of US libraries, which currently serve as models in librarianship.

Implicitly, also a more theoretical aspect may explain the BSB's interest in co-ordinating this particular project, apart from its role as one of the founding member libraries of CERL. There is a stream of consciousness in Germany regretting the fact that our cataloguing rules are not called AACR 2 and that our exchange format has no - MARC suffix, and that therefore we seem to be cut off from the world of international co-operation. The results of the project show that many more differentiations and careful distinctions have to be made in this kind of discussion and that we are not further away from international co-operation and data exchange than any other library community. The awareness of cross border co-operation in cataloguing has a long tradition. It is up to us to translate this well - founded principle into the reality of today which thanks to technology gives us tremendous chances to implement advantageous realizations of co-operation.

In the course of the project we have been able to analyse in total more than 250.000 records for early books coming from six different national sources: Croatia, France, Germany, Italy, Portugal and Sweden; four of these files using UNIMARC as an international exchange format, two of them, Portugal and Croatia, using it as native format. All these files are also meant to be included into the HPB-database, so that one of the results of the project is a practical benefit for a more correct UNIMARC conversion of some of these files, an effort in which the participants as file holders had to invest quite a lot of time and effort.

Participants in the EC project were the Bayerische Staatsbibliothek, Germany (BSB), the Bibliothèque Nationale de France (BNF), the Koninklijke Bibliotheek - Bibliothèque Royale Albert Ier, Belgium (KKB), the Kungliga Biblioteket, Sweden (KBS), the Istituto Centrale per il Catalogo Unico, Roma (ICCU), the British Library, United Kingdom (BL). CERL is closely involved in the project with its Advisory Task Group and the secretary support gratefully offered.

The analysis was based on two approaches, which are necessarily complementary in evaluating formats for bibliographic records, and which only make sense if they are closely interconnected and interacting.

Being a cataloguer, I would say that the first approach is the intellectual analysis of the files, comparing their bibliographic contents, the application of cataloguing rules, all this being translated and reflected in the original format, which is then converted into UNIMARC.

The second approach is the statistical analysis, once the files are in the same format - UNIMARC. For this a software package was produced which allows the statistical analysis of the use of UNIMARC fields and subfields (giving the number of occurrencies, the maximum, minimum and average lengths of fields and subfields) and which also provides a statistical overview of the characters used in these files. The results of this analytical tool, presented in a spreadsheet for each file and in cross-comparison of all files is a powerful instrument especially when combined with the results of the intellectual bibliographical analysis. Its special value is to document

- the areas where there are errors in the file giving invalid UNIMARC data,
- the areas where further investigation is extremely necessary, because there is an important divergence among files,
- other areas where different applications of the format exist but can be handled without harming retrieval,

- and again others, where the awareness of differences can lead to new, commonly agreed standards.

I would hope that everyone can read the report in more detail. Especially by going into the statistics one will probably find new areas of interest and comparison, and maybe sometimes they can even help to decide on questions concerning national format application, UNIMARC conversion, character set use.

The participants agreed that the software developed for the project will be available on the web as shareware. It will be announced on CERL's homepages.

Unfortunately, it is impossible to summarise the whole of the report in just a few minutes. This is a selection from the results which are particularly interesting. All statements are based on the analysis of UNIMARC for monographic material, although - as our files were for early books - more fields and details tend to be used and there is a much smaller conformity in the bibliographic description than for current material.

1. UNIMARC, structured to accommodate the bibliographic description of all kinds of materials formulated according to ISBD principles, is as it stands hospitable to all sorts of original formats, even those where ISBD or ISO 2709 principles are not applied. As such it is a valuable export format where no information contained in the original format gets lost. To integrate files from national formats into a common database, UNIMARC must be supplemented by common agreements, which means a particular format specification for a concrete co-operative project. The very detailed format design of UNIMARC, permitting 166 fields and giving multiple possibilities for locally defined fields, needs particularly careful intellectual monitoring of the consistent application of the format. Converting data into UNIMARC for a common database must follow a subset specification and there must be consequent guidance for each file.

   This specification profits from the application of the software analysis. It shows where there is need to allow for more details in the specification in order not to loose valuable information, and where the specification can be extremely tight, without anybody risking losing data or data definitions.

   The allowance of UNIMARC for local fields is to be cut down in co-operative ventures, those which are maintained must be commonly agreed. Users of UNIMARC should be warned when defining too many local fields. They thus risk being outside the commonly agreed standard. The intellectual analysis of local fields may show that the information could as well be transported in an existing field or subfield, or is of little or no use in a combined database and should be eliminated.

   CERL's experience with its specification was that only two fields were needed for HPB, which in UNIMARC necessarily remain local fields, whereas others, like the fingerprint (012), cataloguer's working notes on sources of information (830), and a field for the title in modern spelling (518) were proposed to the PUC to become standard UNIMARC fields.

The remaining local fields are:

a) alternative forms for names (790, 791, 792), not existing in the UNIMARC format, because an UNIMARC authority format exists, but still necessary because not all bibliographic records are already based on authority files or, even if they are, the link to a local or national authority file makes no sense in a merged database whereas additional searchable access points under the alternative forms for names may give the user some guidance for more exhaustive retrieval.

b) fields for holding or location information, not existing, because a UNIMARC holdings format has not yet been specified, UNIMARC being a format for bibliographic record exchange. In a co-operative database however the need to indicate the locations becomes immediately paramount.

There may be an option to rediscuss both these features and to integrate them into UNIMARC to give guidance for the use of UNIMARC for co-operative database projects of this kind.

2. Although UNIMARC is such a detailed format, the software analysis of the files showed that a comparatively low number of 75 fields is actually used: the maximum being 50 (for Croatia) the minimum 20 (for Sweden); the average 35 - for the others. Both excesses can be explained: Croatia assuring book-in-hand cataloguing for very few items (ca. 2000) applying the full UNIMARC for antiquarian material as original format; Sweden's 18th century bibliographic data being converted into UNIMARC from a complete outsider's broadly defined format (note that this is possible).

This observation allows for some conclusions:

a) The smaller (or more particular) a file is, the more cataloguers and formats tend to go into detail. Careful differentiation of the record tagging takes more time than a broad format application to first provide and then to remember detailed definitions and more field names, to fill them with the appropriate fields contents. It is more inclined to produce mistakes (in the definition, in tagging, in not commonly agreed use, in casual decisions, which are by nature differing).

b) The reason for differentiation and its benefits are usually seen on the retrieval side allowing for the precise indexing and retrieval of special fields. That may always be true in a local environment; for co-operative databases this assumption has to be modified. If a field is not commonly applied, its particular retrieval value is restricted to those items which carry this kind of information. Is it sensible in co-operative database to build up a particular index to retrieve a particular information, which is only given by a small subset of records? It is more likely to index groups of fields integrating particular features into a larger context.

c) A similar reflection applies to data exchange: those who use a very detailed format are not likely to get that degree of detail from others. They therefore have to invest in re-tagging of exchanged data and in supplementing detailed information by editing the record, or they lose the consistency of their database. It is much easier to take detailed data into a broader environment. That can be done by machine procedure, cutting away information which is not wanted or integrating it into more broadly defined fields.

d) Broad and detailed format applications are not recognisable by themselves. The statistics give a good basis for a format specification clearly stating which fields are necessary even in a broad use of the format. From the 75 fields used in our analysis, 28 fields are only used in one file, 13 only in two files. If we exclude the tagging errors appearing in a unique use, the inclusion of subject cataloguing information, the inclusion of serials or of coded fields for modern material, the remaining individual use identifies those libraries which use UNIMARC as their native format, Lisbon and Zagreb. It seems as if already the translation from another format into UNIMARC is a guarantee for a broader use.

e) 12 fields are commonly used in every file. This is the "spine" of the records for books. If we remove the "technical fields" (001, 100, 801), we are left with 101, 200, 210, 215, 300, 500, 700, 701, 702. However, only five fields appear on really every record; - "the technical fields" above and 101 and 200. These fields give clear guidance for grouped indexing. They also define what I would call "umbrella fields", able to take in information which can also be included in a number of more detailed fields. This is easy to show on the notes fields. Libraries can opt for a detailed definition of notes fields using the entire range of 300 fields, or they can input all notes information into one broad field, 300. Indexing of the notes fields could comprise all this range of fields in one common index. Once this minimal standard is carefully designed and agreed, this definition will help libraries in deciding their own local or national format application.

3. As a result of the intellectual analysis as well as the software application one important group of records was identified, where format divergences and cataloguing differences affect the structure of records, of the format and finally the database. This is the problem of multivolume works, a subject dear to cataloguers and format specialists. The EC might initiate another workshop on this particular subject. Germany may learn that they are not the only country to apply multilevel structure. In the files which we analysed only France and Portugal do not apply the linking structure, and that does also explain why their records are in average longer than those of the four others. A common agreement must be found for multivolume works. UNIMARC which for the time being allows for three options can provide more guidance in this field. It is unlikely that one of the two structural approaches - the linking structure or the single record - can be abandoned, as the reasons for both of them are perfectly acceptable, as masses of data are concerned and as the integration into local OPACs or even circulation systems has been achieved. What needs to be agreed on is a translation of the linking structure into UNIMARC which allows for consistent retrieval with data structured on one level for the same item. The intermediate solution may be close to the third option defined in UNIMARC which in our files is not applied by anybody and which foresees a single, complete record for each volume, and where the information common to all volumes is repeated in each record. In any way for retrieval purposes the linking of the lower records to the higher level record must not be by record number only, as this needs systems for retrieval which can handle this kind of linking. Hopefully by more detailed analysis a solution can be found which makes the exchange of differently structured data a good deal easier. That would be a huge step forward towards practical records exchange and cooperation among libraries.

4. The software analysis also comprises the subfields. Although mostly in the descriptive parts of the UNIMARC record, only the $a subfield is mandatory, that does not imply that the broad application of the format is limited to the use of $a. The subfields must be analysed field by field. A smaller problem - without any impact on retrieval, but probably on the exchange of data, seems to be punctuation. Here again, UNIMARC can give clearer guidance. The punctuation must be omitted if translated into subfields. Our actual analysis shows that there is quite a range of options in the application punctuation and subfielding. This can easily be eliminated to assure consistent and reliable use.

5. Statistics show a wide use of coded data in the record label as well as in the coded data fields. Coded data becomes more and more important for retrieval in big databases; they allow sophisticated searching and are language independent. To be useful for this purpose, they need a careful, commonly agreed definition of contents and proposed use, going from "eagerly recommended" to "local", allowing for the integration of narrower definitions into broader ones. What does not help is to define a field like language 101 as mandatory if it has to be filled afterwards by words like "undefined". That is just a means for dealing with data we cannot change, it is not an approach we can adopt for the future.

6. **Character set**

Maybe character sets are even more important for data administration in a common database and for exchange purposes than the use of fields. ISO standards exist, and UNIMARC uses them, but where are we in practice? We don't have to consider A→ Z; 0 → 9; we can leave aside all characters being on an individual position in printing, where agreement for the exchange of data or for the integration in a common database can be reached, for example on a common set of quotation marks, on the hamzah or apostrophe. This kind of agreement is similar to the question of broad or detailed format application, it is to choose between a broad or detailed character set application. The same applies to the use of those characters where the filing is no problem, for example ß (only applied by BSB). Problems will arise in data exchange when the receiving or giving database does not know about these characters beforehand. A clear definition of the characters actually used in the file, as provided by our software package, helps enormously to identify areas where substitution or reediting must be agreed on, the use or non-use of characters must be known in data-exchange. A special problem arises from those characters where filing, indexing and retrieving may be concerned, Umlaute (applied by Sweden, Germany, and Italy), ø (only applied by Sweden). If they are not consistently applied, as both the statistics and intellectual analysis show, the retrieval suffers. Common European agreements for these letters are paramount, they may even lead to double indexing as done for HPB.

7. **Non-sorting-beginning (NSB) and Non-sorting-end (NSE):**

In the control sequence of the UNIMARC character set, the NSB and NSE are defined. They create unnecessary additional problems in European records exchange and database building, although being less a point for format harmonisation than for cataloguing rules. We must agree what we use NSB and NSE for and in which fields they are used. UNIMARC seems to allow for them only in title and notes fields, although it was no problem for Germany to import them into the names fields. It must be resolved whether that is a mandatory feature of the format and in which

fields they have to be used, or whether it is a speciality for more detailed purposes - like filing in a microfiche catalogue. The software analysis shows the use in four files. But the frequency between France, Portugal (only NSB) and Croatia making relatively modest use of them and Germany highly differs. In database environments the role of sorting has to be revised, for indexing other methods exist, although it has to be pointed out that even the identification of articles in the beginning of the title field is largely language dependent.

8. In international data exchange, in the building of common databases, the consistency and compatibility of the format will always solve only one side of the problem. On the other side we have the cataloguing rules and traditions, which may in their differences lead to different format applications, to differing use of fields and subfields, and there are also areas where existing differences are not mirrored in the format but may create problems which are even more difficult to solve. By the intellectual analysis of the records most areas were easily detected: In all areas which are standardised, whether controlled by an authority file or through other consistency methods, standardisation leads to fundamental and substantial differences, which can only be handled and recognised by intellectual input. This is particularly true for names of persons (although here the application of the Copenhagen principles might have solved a lot of discrepancies) and corporate bodies, present in every file. Here the problems are in the area of standardised place and publisher names, which are not consistently used throughout the files, problems are involved in both the allocation of the correct UNIMARC field and the form of names. Place and publisher names are usually designed according to national use, that means that different rules apply to the structuring of the standardised form.

Unfortunately the same is true for the title. Although all files have a 200 field, its content varies from the transcription of the title page as a whole (Swedish 18th century bibliography) to a German understanding manipulating the title saying "Werke" instead of "Goethes Werke". The separation between title and subtitle, title and author statement, the handling of more than one title on the same title page, all this is explained in the national cataloguing rules or even regional or local rule applications, and it is not surprising that it leads to divergence. Differences appear in the use and the designing of uniform titles, collective uniform titles, in the choice of other titles. In those parts of the record where the language of cataloguing intervenes, such as note fields, the differences in description are superseded even by language differences.

These differences must be carefully watched and judged. Not all of them are necessarily hindering co-operation, as we are all more and more acquainted with mixed databases from our own traditions and procedures. It is unlikely that all European libraries catalogue according to the same rules, because of the material, which often enough two cataloguers in the same library would describe differently, because of our differences in languages and catalogue traditions. Nevertheless we have to assure that our records are acceptable and useful in mixed databases. Most of that work has to go into authority control in order to assure consistent retrieval. But there may also be a point in thinking about how to make cataloguing easier, keeping close to the book and inventing less sophisticated rules for cutting or abbreviating the bibliographic description, allocating collective uniform titles, differentiating between all sorts of other titles, and designing them by beautiful German, Italian, Swedish or Croatian - in any case elsewhere, even from our users in their own language unintelligible - names in the footnote.

I think a lot of further work might emerge from the results of our report, each one of them bringing the European libraries closer together, making their record exchange and database building more consistent and easier.

Bearing all this in mind we should not forget that the conformity which is so vital for any co-operation is only one aspect. It is the general agreement, in which I believe completely, to which we all can come. But beneath and beside it, there are the national, regional and local divergences. There is no reason to abandon them, we must in fact take care not to loose this more detailed, more precise information under the broad umbrella, so that it can be of utmost help and use for the entire library community.

Title : AUTHOR : towards a European network for name authority data

UNIMARC Workshop - Luxembourg, Friday 13th September 1996

Author(s): Françoise BOURDON and Sonia ZILLHARDT, Bibliothèque nationale de France

The international exchange of authority records is a subject which has frequently been raised at international conferences, in professional publications and in standards for several years (see references at the end of this paper), but the European Project AUTHOR, started in March 1995, is the first concrete attempt to carry out the plan with 5 national libraries which manage automated authority files with different cataloguing rules, formats and languages.

1. WHY AND HOW THIS PROJECT APPEARED ?

The development of automated national bibliographies, the creation of large national and international pools of bibliographic information in formats which are becoming ever more simple to consult, have all given an enormous boost to the international market for bibliographic records. As they circulate, these bibliographic records carry with them their author access points which are themselves increasingly managed by automated authority files. It would seem logical, therefore, to want to re-use authority data in the same way we re-use bibliographical information. We particularly want to re-use them as the precise identification of an author, personal name or corporate body, requiring a certain type of information which is locally available, where the author works.

According to universal bibliographic control, each national bibliographic agency should establish the authoritative form of a name for its country's authors, both personal and corporate, and for foreign authors, should re-use the authoritative forms established by the agencies of the countries they are from. These principles are difficult to put into practice for several reasons, the main one being that not all agencies manage an authority file, and the second being the difficulty for a given agency to consult the authority files managed by other agencies. Project AUTHOR developed from the need expressed by national bibliographic agencies to have access to the existing authority files throughout the world to re-use the work already done for identifying authors.

The Project is part of the Forum CoBRA's activities (CoBRA = Computerised Bibliographic Record Actions). CoBRA is a concerted action financed by the Libraries Programme of the Directorate General XIII of the Commission of the European Communities. Started in 1993, Cobra aims to develop the participation of national libraries in research and development programmes. Project AUTHOR is the result of a direct partnership between the European national libraries and DGXIII. The European Commission finances 100 per cent of certain costs, and the budget comes to 155 000 ECU.

It is necessary to recall that the projects issued from CoBRA partly derived from the European Project of a unique CD-ROM for several official national bibliographies published by different national libraries (LIBACT1/CDBIB, 1989-1992). This CD-ROM proved the feasibility of international cooperation for exchanging bibliographic data recorded with different formats, different cataloguing rules, and different languages. Project AUTHOR is a successor of this previous Project and aims to implement its recommendations.

1 This paper was originally presented at the 62nd IFLA General conference (25-31 August 1996, Beijing) to the Division on bibliographic control, Section on bibliography

The partners of the Project AUTHOR are the following :

- Bibliothèque Royale Albert 1er (Belgium),
- Biblioteca Nacional (Spain),
- The British Library (UK),
- Instituto da Biblioteca Nacional e do Livro (Portugal),
- Bibliothèque nationale de France, which is the scientific and administrative coordinator of the Project.

2. OBJECTIVES EXPECTED FROM PROJECT AUTHOR

2.1. Studying the technical feasibility of the following points

- to give access to authority files for names of persons and corporate bodies, at the international level by means of a test bed platform and to define a target technical architecture;
- to convert authority data produced by the national libraries in the Project to the international exchange format for authority data prepared by IFLA: UNIMARC/Authorities
- to re-use authority data made available in this way in the current practice of cataloguing.

2.2. Implementing and promoting results of previous European Projects

- Project UseMARCON (User-Controlled Generic MARC Converter)

  This Project managed by the Koninklijke Bibliotheek (Netherlands) aims to develop a Generic MARC Convertor, that is to say a software which allows a librarian to state himself/herself the conversion rules necessary to convert bibliographic records from any source MARC format into another target MARC format. This "toolbox" uses UNIMARC as a pivot format. At the present time, the software is tested and the Project should be ended this autumn. Two of the partners of the UseMARCON Project are also partners of the AUTHOR Project (The British Library,UK - and the Instituto da Biblioteca Nacional e do Livro, Portugal) and this situation should make the exploitation of the UseMARCON results by the Project AUTHOR easier.

- Project EUROPAGATE

  Ended in 1995 this Project developed a gateway between a Z39.50 client and a ISO SR server, and vice versa between a Z39.50 server and a ISO SR client, in order to give remote access to bibliographic databases. It solved also the technical problems raised by access to multiple servers each of them having its own characteristics. This software, easily portable, offers a standardised interface between servers which give access to bibliographic databases and largely facilitates international connections. Project AUTHOR will take EUROPAGATE into account in searching for a technical architecture.

2.3. Expected results are the following :

- to elaborate conversion tables for the partners' authority files, from their national format to the UNIMARC/Authorities format ; these tables should be re-usable later on;

- to examine problems raised by the elaboration of conversion tables and to propose recommendations to the IFLA UNIMARC Permanent Committee to have UNIMARC updated according to the requirements;
- to give access to authority data through the Z39.50 protocol and the WEB;
- to propose to IFLA a definition of the minimal content of an authority record intended for international exchange, in close relation with the IFLA UBCIM Working Group created in May 1996 to work on the same topic and of which The Bibliothèque nationale de France and The British Library are also members;
- to infer a target technical architecture able to be opened to other libraries from the test bed platform.

3. WORKPLAN AND CALENDAR

The work falls into 3 main phases :

- background, preliminary study and technical study;
- development of the test bed;
- test evaluations and recommendations.

Initially, the Project was going on for 12 months (March 1995-March 1996). Partners asked the Commision for an extension of the Project continuance in order to be able to test the UseMARCON software of which the beta version was expected in July 1996.

From the moment the partners chose a technical approach based on access to networked databases, delays appeared in the definition of the technical architecture.

3.1. Preparation of conversion tables : April-June 1995

Since the use of the UNIMARC/Authorities format is part of the Project objectives, elaboration of conversion tables was made right away. Of course, the choice of a technical scenario for the test-bed platform will determine how these tables will be used in the framework of the Project.

After the partners came to an agreement concerning a standard conversion table model, each national library prepared its conversion table from its national format to UNIMARC/Authorities, except for Portugal which already worked in UNIMARC. Portugal co-ordinated this work and kept an eye on the coherence of results : common data elements coded in different ways in national formats must be put in the same UNIMARC field or subfield after conversion ; but, information of different types keyed by the same way in the different national formats must be managed differently during conversion in order to give non ambiguous results in UNIMARC.

3.2. Background study and preliminary study

To carry out the Project, the partners appointed a technical consultant: Bureau Van Dijk. The consultant visited each participating library to study its authority file in its original environment and to record the expectations of the partners in the framework of the Project. A report was delivered at the end of 1995 from which it appeared that the Project would have to take into account:

- 5 cataloguing languages: English, Spanish, French, Dutch and Portuguese, and a bilingual catalogue French/Dutch;
- 5 cataloguing rules: AACR2 (UK) and 4 different national standards for Spain, France, Belgium and Portugal;

- 5 MARC formats: IBERMARC (Spain), INTERMARC (France), BLMARC (UK), UNIMARC (Portugal) and KBRMARC (Belgium);
- 4 softwares: ARIADNA (Spain), GEAC (France and Portugal), VUBIS (Belgium) and WLN (UK).

Partners defined their need as follows:

- to search the authority data on line (in preference to a CD-ROM) to have access to up to date information;
- to display records in UNIMARC;
- to re-use data by copying and re-keyeing relevant information in the local file, and not by automatic downloading and uploading records: because this uploading would require conversion tables from the UNIMARC/Authorities format to each national format. The feasibility of such conversions is not yet established, and because of the different cataloguing rules which are in competition, imported records should be edited anyway before being integrated in the local file.

3.3. Choice of a technical scenario (June 1996):

The technical study started in August 1995 during a meeting with the consultant and the BNF pilots of the Project. A first list of possible scenarios for the test bed platform was delivered by the consultant in October 1995: 6 scenarios were proposed together with a statement of advantages and disadvantages of each.

The choice of a technical scenario depends on the combination of 3 criteria:

- data format: unique format (i.e UNIMARC format, in a unique or in separate servers) or separate formats (i.e. the different national formats);
- file structure: unique file (that implies a conversion of the different national formats to UNIMARC) or separate files (that makes optional the conversion of the national formats to UNIMARC);
- server structure: unique server (the management of which must be assumed by a National Library or a commercial company) or separate servers (which are a priori the servers of the different National Libraries).

3.3.1 The target system

The preferred scenario for the target system is the scenario which allows remote access to the name authority files of the different national libraries. Access to the distributed databases is through the Z39.50 protocol implemented on each server of a name authority file. This permits a unique request to be sent to different severs and to obtain a global answer giving the results of the search.

* Each server :

- is updated and managed by the library;
- gives access to the records as a single batch.

* The conversion to UNIMARC will be done in real time

3.3.2. The prototype

In order to prove the feasibility of the target system, a prototype will be built according to the following technical and functional specifications :

- a system just for testing access to the records, and by that very fact, with a short life;
- a prototype at a cheap rate;
- few authority records: each partner will give a sample defined according to common criteria;
- a unique sever independent of existing systems in the libraries;
- a previous conversion of the sample records made thanks to the UseMARCON software (and not in real time as designed in the target system);
- direct access via INTERNET by the means of a WEB navigator or a Z39.50 client. The Z39.50/WEB gateway allows the user having a WEB navigator to receive HTML pages. On these pages the user can select one or several databases and make his/her request. Then the request is translated into Z39.50 and passed on the server. The same process can be applied to the answer;
- the prototype will be built as a unique database (OPAA = Open Public Authority Access) which will simulate the access to the 5 databases of the national libraries;
- the fields 2XX, 4XX and 5XX of the UNIMARC/Authorities format will be searched.

4. NEXT STEPS

4.1 To test the use of UseMARCON to convert MARC authority formats

The beta version of UseMARCON will have to be tested to verify if the software prepared for converting MARC bibliographic formats can also be used to convert MARC authority formats. To do that, partners will use the conversion tables prepared from their national authority format to UNIMARC/Authorities. What will make this work easy is the fact that 2 partners of the AUTHOR Project are also partners of the UseMARCON Project: Portugal and UK. From what these two partners inferred from their preliminary approach of the specific aspects of authority formats, no major difficulties should be faced. The main problem could arise from the necessity to generate a record (authority record or reference record) according to the target format (UNIMARC) from a tag (parallel heading or cross reference) of an authority record written according to the source format (especially INTERMARC). Some developments of UseMARCON could be possibly asked to make this software a real universal convertor of the MARC formats both for bibliographic and authority records.

At the end of this period, each partner should be able to convert its sample of authority records to be loaded on the test prototype using UseMARCON.

4.2. To integrate softwares and data with the prototype. A preliminary inquiry with some different European companies assured us of the technical feasibility of the solution retained for the prototype architecture in the framework of the Project. The prototype should be located on the pilot site, that is to say at the Bibliothèque nationale de France.

4.3. To test and evaluate the re-use of authority data by cataloguers

Testings should take place during the second quarter of 1997. An evaluation guide will be circulated to the participants, but right now the way the tests will be carried out has not been decided: perhaps there will be a questionnaire, perhaps a case study, etc.? While the conceptual work was intentionally limited to the 5 partners, tests could be done on a larger scale. More European libraries and institutions will be invited to participate.

So today we are on the road to success considering the international cooperation of authority data: the current work between the Library of Congress and the British Library to develop an authority file common to these two institutions and the implementation of Project AUTHOR give evidence of it. We are testing what was but simply a dream for many years: the exchange of authority data at the international level. Technical developments make the concrete realisation easy: librarians adapt themselves to increasingly sophisticated communication tools perfected outside their field of action (INTERNET, EUROPAGATE, etc), but also contribute directly to perfect the tools they need (UseMARCON for example), with the result that, far from deleting differences between national cataloguing practices, these new tools allow us to manage them and urge us to take an advantage of the richness of our neighbours without being anxious to loose our distinctions. It remains for us to learn how to exploit our differences to make our catalogues more complete without having to duplicate work and so to save a lot of time and a lot of staff.

OCLC Report

UNIMARC Workshop - Luxembourg, Friday 13th September 1996

OCLC UNIMARC Development - Status Report

Introduction

OCLC's decision to develop a UNIMARC capability can be traced to 2 events which occurred in February 1995:

i) The OCLC Board of Trustees approved a number of product enhancements proposed by the international directors of OCLC in order to support OCLC's international growth. One of these enhancements was the development of a UNIMARC capability.

ii) OCLC entered into an agreement with the National Library of the Czech Republic to load the Czech National Bibliography into the OCLC Online Union Catalogue. The agreement specified that the records should be delivered in UNIMARC format.

It should be noted that unlike the other projects described during this workshop the OCLC Development is not a research project but a production facility.

Background

OCLC is an international cooperative providing services to libraries in 64 countries. Last month, August 1996, OCLC celebrated the 25th anniversary of the OCLC Online Union Catalog (OLUC), its shared bibliographic resource containing more than 35 million bibliographic records and 600 million holdings locations.

The OLUC is maintained in USMARC format and the records in the database conform to the Anglo-American Cataloguing Rules, 2nd edition (AACR2) and the Library of Congress Name Authority File. The records are derived from three main sources:

i) national libraries including Library of Congress, British Library, National Library of Canada and National Library of Australia and National Library of Czech Republic.

ii) current cataloguing performed by member libraries

and

iii) retrospective cataloguing performed online by member libraries and the OCLC Retrospective Conversion Unit.

The OLUC is the foundation for OCLC services including online cataloguing, CD based cataloguing products, inter-library loan, retrospective conversion, and reference services.

OCLC's experience of format conversion options

The enrichment of the OLUC is a high priority for OCLC. As OCLC has expanded into new countries and regions there has been the requirement both to import and export records in formats other than USMARC. In the past 10-15 years this problem has been addressed by the use of 3rd parties.

- Library of Congress has developed conversion software which it uses for converting national library files. LC has made this software available to OCLC on a case-by-case basis - for example it is used in converting the British Library's UKMARC files - but LC is reticent to become a large scale format converter.

- CURL - Consortium of University and Research Libraries based in UK - has developed UKMARC-USMARC and USMARC-UKMARC conversion software which it uses to exchange bibliographic records with OCLC - both in contributing records and holdings to OCLC and receiving records on behalf of its members.

- commercial 3rd party vendors who provided customized conversion software for individual libraries on a contract basis. These services are usually used for large retrospective conversion agreements. The advantage is that the service can be tailored exactly to the library's individual requirements (including local data) but the disadvantages are both in timeliness (tapes/disks transferred by mail) and costs which can be prohibitive for small, frequent conversion needed for current cataloguing.

- local system vendors who provide interfaces to OCLC which involve format conversion as part of the service to particular library communities such as SLS (UK, Spain, Sweden) and EOSI (TinLib) primarily in Central and Eastern Europe.

As these services proliferate they become difficult to manage and maintain. USMARC changes need to be reflected in changes to format conversion software and disseminating and coordinating such changes becomes more and more difficult when working with 3rd parties who have differing priorities and timeframes.

The decision to support UNIMARC

OCLC's experience of format conversion led it to conclude that it needed to support a production facility for the exchange of bibliographic records internationally which did not rely on libraries adopting a 'foreign' national MARC format.

The decision to support UNIMARC was made on the following grounds:

i) The format had published documentation and an established maintenance organization

ii) Many national libraries were adopting UNIMARC as a primary or secondary format for exchange

iii) Many libraries in Central and Eastern Europe were adopting UNIMARC.

The importance of a UNIMARC capability was emphasized in the OCLC International Business Plan which was approved by OCLC's Board of Trustees in February 1995 "to develop or to acquire software that converts UNIMARC bibliographic records to USMARC and vice versa". In the same month OCLC concluded an agreement with the National Library of the Czech Republic to load the Czech National Bibliography into OLUC. The records would be delivered in UNIMARC format.

Four applications of the UNIMARC/USMARC Conversion are foreseen, to be accomplished in phases: UNIMARC output through PRISM export; UNIMARC output through subscription and tape services; UNIMARC batchload capability for the OLUC; UNIMARC output from CatCD for Windows and other micro cataloguing products. Output of UNIMARC records would also be supported for the various Retrospective Conversion options, including RETROCON and MICROCON. These UNIMARC applications can be developed independently and will be market-ready at different times.

All future USMARC maintenance projects at OCLC will need to address the UNIMARC/USMARC conversion. UNIMARC output will eventually need to be available across the range of OCLC services and output products, built into any new product developments and folded into existing products via enhancements. UNIMARC output for authority records is not currently planned.

OCLC's conversion efforts are based on the 1994 UNIMARC Manual Bibliographic Format, 2nd edition, which included improved provisions for component parts, microforms, and three-dimensional artifacts and realia.

Challenges

To become more familiar with the format as it is actually implemented in real situations, members of the OCLC UNIMARC group met in March and April of 1996 with representatives of three European national libraries: the National Library of the Czech Republic, Prague; the Russian National Public Library for Science and Technology, Moscow; and the National and University Library, Zagreb, Croatia. Among the questions discussed were:

- Given the imperfection of all conversion processes, what level of data loss will be tolerable?

- Since UNIMARC allows considerable variation in local implementation, what sorts of differences can we expect to find from institution to institution?

- How can we best deal with such advanced UNIMARC practices as embedded fields and sophisticated linking of records (both bibliographic and authority) that USMARC either cannot currently handle or deals with in a much more rudimentary fashion?

- What impact do different cataloguing rules have on individual implementations of UNIMARC?

Unlike such past conversions undertaken by OCLC as that between UKMARC and USMARC, the UNIMARC conversion involves a format developed from a significantly different perspective and seemingly more distant from the catalog card-inspired derivatives of USMARC. To mention just a few of the many differences:

Character sets are different, ALA for USMARC and various ISO sets (2022, 646, 6630, etc.) for UNIMARC, with most diacritics and special characters needing translation.

UNIMARC de-emphasizes the notion of "main entry" by placing all intellectual responsibility in its 7-- area. USMARC explicitly separates main entries in the 1XX area.

UNIMARC raises Family Name entries to a status equivalent to that of personal, corporate, and uniform title headings; in AACR2 and hence in USMARC, family names are rare and designated by an indicator in the Personal Name field.

Whereas USMARC dumps most general notes into 500 fields, UNIMARC divides its "general" notes into over a dozen different fields corresponding to other areas of the bibliographic record.

UNIMARC provides for embedded fields, with possibilities for complex hierarchical links both among bibliographic records and between authority and bibliographic records.

As a result of the vast differences between the formats, virtually every field and most subfields within each field must undergo conversion. This includes character set translations, tag renamings, subfield reassignments, indicator modifications, and in numerous cases, manipulation of data within certain fields and subfields.

Because of the complexity of the conversions in each direction, there will be an unavoidable loss of some data. In some cases where there was no logically apparent equivalent field in the other format, OCLC chose to embed the original field wholesale in the USMARC 866 field (Foreign MARC Information Field) or in a locally-assigned UNIMARC 9-- field.

Current Status

Subfield-by-subfield draft specifications for the conversion in both directions between UNIMARC and USMARC were begun in June 1995 and completed in January 1996. At that time, programmers began coding and testing the UNIMARC to USMARC conversion. That effort was substantially completed in July 1996 when OCLC sent a file of converted records to the National Library of the Czech Republic for their judgement. Since then, the National Library has been studying the resulting records and correcting certain data problems at their end. OCLC has been fine-tuning the software at our end. When all conversion problems have been worked out to each party's satisfaction, OCLC will be able to run these Czech records through its Batchload software to match, set holdings, and load.

Meanwhile, coding of the USMARC to UNIMARC conversion software is well underway, with completion scheduled for June 1997. The software has been designed to be able to run on multiple hardware platforms.

The success of the project to date has been dependent on international cooperation on a number of levels. Firstly, OCLC staff whilst experts in USMARC had little experience of UNIMARC. The close cooperation with the National Library of the Czech Republic and the input from other libraries enabled OCLC to gain a practical insight into the use of the format.

OCLC was also able to gain a good impression on the likely variation in interpretation and application of the UNIMARC format through the receipt of test files from other libraries including ICCU, Italy; Deutsche Bibliothek and the National Library of Portugal. Whilst initially surprised at the difficulty in gaining test files the availability of such data was invaluable in testing our conversion software!