

Review Document: Executive Report on the CURL MARC21 Database Meetings

Introduction

Between September and November 02 three regional meetings were convened by CURL and MIMAS to canvas ideas on the form and functionality of the new MARC21 database. These meetings were held in Manchester, London and Edinburgh, and were attended by cataloguing and systems staff from 25 out of 27 CURL Member, Associate and Partner institutions. Each meeting was also attended by Mike Mertens (CURL Database Officer), Sarah Davnall (COPAC programmer, MIMAS), and Nicholas Syrotiuk (MARC21 Officer, MIMAS).

A set of focus questions (see Appendix 1) was composed co-operatively by CURL and MIMAS staff, and disseminated to delegates before each meeting. In addition, after the first meeting in Manchester, delegates at subsequent meetings were also given a set of summary documents from previous meetings.

What follows is an executive summary of the three documents produced after each of the three meetings. The below consists therefore of a series of recommendations for the specification, based on the ideas put forward by delegates. The document will concentrate on the issues raised by delegates in the following order:

- **Search parameters and associated indexes.**
- **New interface.**
- **Record representation**
- **Authority data**
- **Non-Roman characters**
- **Z39.50 and the Bath Profile**

Unless otherwise specified, the field codes used in the document are the ones available in MARC21.

1) Search parameters and associated indexes.

Additional indexes.

The following additional indexes were requested:

Search type	Associated fields (MARC21)
Statement of Scale for maps	034
Opus or thematic catalogue number (printed music and sound recordings)	240 \$n, 700 \$n
Publisher number (sound recordings, printed music, Videorecordings.)	028
Uniform Resource Identifier	856 \$u
Fingerprint identifier ¹	026

¹ This is a relatively new field, defined in the web-based version of the format. It is used to assist in the identification of antiquarian books by recording information comprising groups of characters taken

The following fields were also suggested for indexing for searching purposes so that search terms could be employed either to limit a search at the outset, that is, applied to the whole session, or used to qualify an existing search result set.

Refinement type	Associated fields
Date range	008, 260
Material type	Leader/06 & 07
Language code	008
Place of publication	008, 260
CURL institution code	003, 035
Microform	007
Bibliographical level	Leader/07

Futhermore, there was a desire to narrow search sets by using a third option, the **ALSO** command, in order to limit further and create subsets from original search result sets by pagination and place of publication (**260 and 300**)

Finally, delegates were keen to be able to browse certain indexes on the new database. This would require the ability of Members' Z39.50 Clients to use the **SCAN** function.

2) New interface

Many suggestions were made as to the function and appearance of any replacement for the current Telnet interface. Concerns were raised about the Telnet interface regarding its lack of security. On the whole, a web interface is to be preferred for reasons of interoperability and familiarity: this is especially pertinent with regard to the fact that much cataloguing is now done by non-professional staff. The requirements of any interface should be as follows:

- User-friendliness
 - Ease of navigation through search results and back to the main search area
 - This could potentially include:
 - the option of sorting a result set by title or by reverse date
 - the navigation within the interface to be improved (e.g., after viewing a record selected from the middle of the result set, the searcher wants to be able to navigate back to the middle of the list)
- Security
- Record selection for later batch downloading, with the ability to select more than one record from the result set.
- Offline batch retrieval facility
- Economy of use
- Precision of use
- The ability to display accented characters

from specified positions on specified pages of the book, in accordance with the principles laid down in various published guidelines. See: <http://lcweb.loc.gov/marc/bibliographic/ecbdunderstand.html#mrcb026>

- Short display to convey coded information about authorised headings in a given bibliographic record

The issue for CURL about the eventual form of the interface is a financial one, for the development of a web-based interface was not part of the original brief given to MIMAS. Work could be done to enhance the current Telnet interface, but an exceptional opportunity exists for CURL to future-proof its services by choosing a web-based product to replace Telnet. Access via a Z39.50 interface would continue, but it is also important strategically for CURL to provide itself with a non-Z39.50 route to its own database for two main reasons:

- Uniformity of access and service: the variability of Z39.50 Clients hinders this
- To make the CURL service independent of LMS vendor development programmes: investment in and upgrading the Z39.50 technology of proprietary LMS is not optimal for CURL purposes.

Subject to cost, the ideal solution would be to emulate the current Eureka web-interface, in the sense that the Eureka offers both web-based **and** command-line text searches, with the latter retaining the syntax of the RLIN database (Telnet version).

3) Record representation

There was a debate on the types of Database Model to be employed. There are three basic current database models: the OCLC model, the CURL model and the RLIN model.

OCLC Model.

This employs a master record structure. Ideally, the OCLC database contains exactly one bibliographic record, called a “master record,” for every bibliographic item that is catalogued by a member library. Attached to each master record is a list of member libraries that hold the item.

CURL model

In this model, the database contains all the bibliographic records contributed by member libraries, and includes many duplicate records for the same bibliographic item.

RLIN model

This groups together bibliographic records for the same bibliographic item into a “cluster” by comparing certain data elements in the records. One record in the cluster is called the Primary Cluster Member (in RLIN terminology) and is used for display while the others are hidden in the background but can be viewed on request.

An additional model used by MIMAS is the **COPAC model**: this employs record consolidation. In this model, bibliographic records for the same bibliographic item are matched, de-duplicated and consolidated into a single record. Of the other types, this model is the most similar to the mode employed by RLG for the RLIN model.

The RLIN model of clustering with a Primary Cluster Member is to be recommended here. This was the majority view at the meetings, and is to be commended for the following reasons:

- It offers a significant compromise between reducing selection time by less qualified staff and the specific requirements of specialist cataloguers, who may need to review many records for a precise match.
- Interoperability. The model is familiar to technical staff, and the clustering model can be readily accommodated by the use of drop-down menus as used in the Eureka interface.

The practical objection raised that the Primary Cluster Member may not represent a full or ideal record for certain Member institutions can, in using this model, be overcome by the method of individual Library Profiles. In this way, different libraries can specify what fields they would require to be present in their ideal version of a Primary Cluster Member at record retrieval time. Therefore, staff with a range of requirements, from copy-cataloguing to cataloguing early printed books, can all be accommodated.

4) Authority Data

The views on the loading or accessibility of authority data through CURL services were mixed. This ambiguity was partly created by potential cost factors, and the feeling that any funds used to buy in or mount authority data might be better spent on other CURL initiatives. Authority data could be made available in several ways:

Data	Advantages	Disadvantages
Web LC Authorities service (Mirrored)	Free	Only in Latin-1 set
	Ease of access	Limit to concurrent users
		Only individual authority record downloads
LC Name and Subject Authorities (via ftp)	Distinct cost benefit for members	Expensive to buy and maintain (initially, ca. 28,000 US Dollars; thereafter 15,000 US Dollars per annum for updates)
	Increased standardization	
	More efficient copy-cataloguing	
	Improved searching (over time)	
Share LC NSA file subscription with the British Library	As above	Much reduced cost: retrospective files cost, plus shared future subscription

There is a great opportunity here, dramatically to increase the use of the database to Members by buying in this data.

There are also several ways in which to use authority data. On the whole, it was felt inadvisable to check the headings of incoming bibliographical records at load time. This is because authority data is very dynamic, and many headings would become outdated before long, having to undergo future global fixes in any case in order to remain current after first load.

The two most useful aspect of access to Authority data were envisaged at the point of download, and/or cataloguing. Checking authority data at the point of download would require a tag to display the authority data status of any record in short display. In the light of the fact that much cataloguing is undertaken by clerical staff under supervision, this would improve copy-cataloguing, and foster the standardization of database content and cataloguing practice.

5) Non-Roman Characters

The question of the representation of non-Roman characters was initially considered to be of a lower priority. The response to the questions on whether staff thought that there would be a demand both to have non-Roman vernacular scripts displayed and used as a search option was mixed and uncertain, although there was some desire that the 880 Alternate Graphic Representation field should be retained where it exists in incoming records

However, this is one area in which CURL needs to think seriously about future-proofing its services, for it does seem likely that demand for searching using non-Roman scripts and having results displayed in non-Roman characters will become significant, if not in the short term, then certainly in the medium term.

In the light of the above, the following recommendations should be made

A) That Unicode be used as the method of storing data for non-Roman characters. This is currently supported by Aleph, with Voyager, Innopac and GEAC Advance moving towards support of Unicode. Unicode is likely to become the standard means of encrypting such data, with future-oriented encoding tools/methods such as XML, Java, ECMAScript (JavaScript), LDAP, CORBA 3.0, WML being compliant. It is also compatible with ISO/IEC 10646 (otherwise known as UCS). UCS is the first officially standardized coded character set with the purpose to eventually include all characters used in all the written languages in the world. UCS is intended to be usable both for internal data representation in computer systems and in data communication. Microsoft, Novell, Apple employ UCS: it is also implemented in open source software like Linux, and is already included in advanced data communication standards like HTML².

B) That CURL sets up and enforces a new part of its cataloguing standard in relation to the 880 field. It appears that there are no restrictions on what should appear in the

² See: <http://www.nada.kth.se/i18n/ucs/unicode-iso10646-oview.html#1>

880 field, but that uniformity of practice in CURL as to which fields are to be indexed for vernacular script will need to come about.

C) That CURL consults closely with RLG on policies relating to non-Roman characters.

It is also important to note that, if anywhere, it will be here that Member library Z39.50 Clients may pose a problem in the full provision of a service that renders non-Roman characters. This is one of a few issues regarding conformity between Members Z39.50 Clients and the target Z39.50 server at MIMAS that will also have to be dealt with.

6) Z39.50 and the Bath Profile

With the introduction of the MARC 21 database, it is planned that the CURL Z39.50 target will be set up to be conformant with Bath Profile functional area A Level 1³. The Bath Profile is a set of internationally agreed minimum attributes for the Z39.50 protocol that ensure a core set of functions and facilities between different systems that use Z39.50 technology. During all these discussions, the ability to provide some of the additional functionality through the indexing of further MARC record fields to Members was questioned on the basis that not all the desired searching parameters had corresponding attributes in the Bath Profile. However, there are local Use attributes in the Bath Profile that allow for customisation, and these can be used to support functionality beyond the core standard attribute set.

Conversely, there is a question relating to the searching in and display of non-Roman characters which concerns the ability of CURL Members' Z39.50 Clients to accommodate this service, if provided by MIMAS. Although moves towards supporting Unicode are being undertaken by several Library Management Systems (LMS) vendors, as things now stand technically, it is not likely that all CURL members would be able to make use of services that employ vernacular scripts. There are two ways in which CURL could bring about the necessary conformity needed here between the Z39.50 Server at MIMAS and Members' Z39.50 Clients.

- Identify a reliable piece of 3rd party software to integrate with Members' LMS
- CURL to act as a body to put pressure on LMS vendors to upgrade their proprietary Z39.50 Clients, and hasten the support of Unicode.

Neither policy is without problems. None of the current 3rd party Z39.50 Client software packages is recommended by MIMAS as being, on present evidence, readily integratable with Members LMS's. On the other hand, opinion and experience on the part of CURL Member systems staff present at the meetings point towards a lack of concern on the part of LMS vendors to improve their embedded Z39.50 Clients. This is largely to do with the market position of the UK in the eyes of LMS vendors whose main customer base is in the United States, where demand for high levels of Z39.50 functionality is low.

³ See: <http://www.nlc-bnc.ca/bath/bp-current.htm>

It is to be recognised however that the efforts of individual Members in gaining recognition of their Z39.50 Client problems have been reportedly in vain. Although it is not certain to succeed, CURL should back up Members, and approach the individual LMS vendors for a positive response. As stated there is some movement in this direction towards the adoption of Unicode by certain vendors, and CURL as a body ought to consider adding to this momentum.

In Conclusion

The above recommendations are made in the light of the consultation with CURL staff, which is ongoing, and in consideration of wider developments regarding the potential branded strength of the CURL database. In terms of the prioritization of the work to be done as originally proposed by MIMAS, the low position occupied by the evaluation of subject and name authority files may need to be revisited in the light of demand. Development work on the user interface should also be brought forward as a greater priority, and given policy and financial assistance from CURL.

Mike Mertens
CURL Database Officer & Deputy Executive Secretary

December, 2002.

Appendix 1: Original Focus Questions

Improving the CURL Database

The questions below indicate the issues upon which we wish to consult you. They are not definitive, but are meant as starting points for discussion.

Interfaces

1. Will Z39.50 be the mechanism that you will be using to access remote databases?
2. What other interfaces are your systems able to support?
- 3 Will you continue to need Telnet access to CURL?
4. Would you like to have a Web interface with the same functions as the Telnet interface?

Searching

1. How can existing search features be improved?
2. What new search facilities would you like?
3. For Z39.50, do you want search facilities to conform to the Bath Profile?
4. Do you want to be able to include accented characters in search terms?
5. Do you want to be able to search in vernacular script?

Downloading

1. Should the batch download (file create) facility be retained?
2. Do any records which you download contain data features which cause local import problems?
3. When downloading records, which character set would you want: MARC8 or Unicode?

General

1. What other facilities could CURL put in place to make Record Retrieval easier or more efficient?
2. If authority data were hosted with the CURL database, how would you like to make use of it?
3. Should the CURL service be made (more) like the other major utilities which you use?