LIBER QUARTERLY, ISSN 1435-5205
© LIBER 2001. All rights reserved
K.G. Saur, Munich. Printed in Germany

Cataloguing in the Digital Age

by STUART EDE

THE ROLE OF CATALOGUING

Some fundamental changes are happening, or will have to happen, to adapt our catalogues to cope with what I call the Digital Age - an age in which we have to meet our users' requirement to find the information they want irrespective of whether it is in print or digital form and whether it is held locally or on the Web. This is posing some huge challenges - indeed it is forcing a radical rethink of what libraries are for and what their role should be. The whole of the information world is at a crossroads.

It is my belief that cataloguing in some form or other is still the key to access - indeed it is even more important if users are to find the knowledge and information they need amongst the tidal wave of raw data. However, cataloguing is an expensive business, and the economic pressures will be made worse by the increase in publishing in both digital and print formats. This will force the even wider re-use of records than is already occurring, but efficient re-use is only possible if records are prepared to common standards. A major focus of this paper is therefore the development of standards and the factors that will influence the direction that takes.

THE RUSH TO STANDARDISE

The Web and digital publishing communities have realised the importance of standards, and there is a flurry of activity to develop them. Unfortunately there is a multiplicity of standards in development, and it can be difficult to find one's way through the maze. This situation begs a number of questions:

- What will emerge from the flurry of standards for digital documents?
- Are we in danger of separating the digital world from the print world through separate standards?
- Will standards encompass both print and digital media?

- What is the future for MARC?
- Much development is happening in the trade sector - how will library and trade sectors interact?

One of the main purposes of this paper is to take a view of the future by attempting to answer these questions and thereby stimulate debate.

OVERVIEW OF STANDARDS

In one way the plethora of standards that is being developed might be viewed as a healthy sign, but on the other hand we do not want a Tower of Babel, particularly if different communities adopt different standards, e.g. trade from libraries, video from books, digital from print etc.

A brief survey of the prominent standards will help set the scene before attempting to look in to the future to answer the questions posed above. Cataloguing standards - leaving aside communications and catalogue search protocols such as Z39.50 - can be categorised as:

- data dictionaries (lists of data elements)
- formats (how the data are carried)
- identifiers (to uniquely identify documents, authors etc)
- cataloguing rules (by which data are presented consistently to facilitate retrieval).

The data dictionaries with greatest significance for libraries are Dublin Core, which originated at a conference hosted by OCLC in Dublin Ohio, and ONIX (Online Information eXchange) which has emerged from the book trade as a subset of the wider EPICS (EDItEUR Product Information Communication Standards) data dictionary. EDItEUR is the international organisation concerned with developing standards to support trade applications, but who are also keen to ensure collaboration with the library sector.

In the formats category we have MARC (MAchine Readable Cataloguing), which has been building up its capability to accommodate records for digital documents. Meanwhile the new kid on the block is XML (eXtensible Markup Language) with its specially designed metadata format, the Resource Description Framework (RDF) for embedding records in XML documents.

Identifiers are important to uniquely identify an item to facilitate the reuse of common components and to minimise data traffic. It is outside the scope of this paper to say much about identifiers, but it is just worth mentioning INTERPARTY, an EU funded project that is about to start. This is an example of a convergence of interest between the library and trade sectors. The library world has through the International Federation of Library Associations (IFLA) been working towards a standard identifier to allow the unique identification of persons and the re-use of accurate name authority data packages. However, this requires name authority registries in each country, and there has not been the resource to achieve that. Meanwhile in the trade sector rights management is a hot topic. Publishers, authors and other rights holders naturally want to derive revenue from the electronic market in their digital documents, and the unique identification of rights holders is essential to support this. The potential profits from this marketplace could mean that the network of registries foreseen by IFLA might be funded by trade organisations. This convergence of purpose is useful to illustrate what may become an increasingly familiar phenomenon.

The final component of the information architecture is the cataloguing codes used to provide the intellectual structure for the data. Without that structure the data loses much of its meaning.

DATA DICTIONARIES

Dublin Core[1] was originally developed for resource discovery of digital documents on the Web. It comprises 15 data elements and is meant to be simple and cheap to apply, especially by the originators of the documents.

There has been much debate about the 15 elements, and whether or not to extend them to meet new requirements. That issue has been met to a degree by the addition of qualifiers to those 15 data elements. The philosophy of Dublin Core development is that additions to the basic standard should be developed by user communities. They will draw up sub-structures beneath the 15 elements to accommodate the special requirements of their applications. However, this creates a problem of keeping different communities' sub-structures consistent and avoiding overlap. There are differing views about this philosophy, and there are those even within the DC community who question whether DC should expand in to new sectors or whether it would be better to concentrate on its strength in the Web resource discovery arena.

Dublin Core is not tied to a specific format, but it can be put into XML. Nor is DC limited to digital documents. A prominent member of the Dublin Core community is convinced Dublin Core will be used for print documents as well - of which more later.

The other major data dictionary of note for libraries is ONIX[2]. It is a subset of the EPICS data dictionary developed by EDItEUR[3] and endorsed by the influential <indecs> project[4]. The ONIX subset emerged from the Association of American Publishers discussions on how to get e-commerce applications under way in the trade sector. It is driven in particular by the needs of on-line booksellers where rich metadata, e.g. reviews, blurbs, extracts and images, is necessary to make up for the inability of customers to browse the books in the flesh.

ONIX is implemented in XML. Version 2.0 has been published this year, which can accommodate multimedia publications and e-books. ONIX is being adopted by an increasing number of trade bodies, for example in the US, UK, Germany and Latin America, and by international bodies like the ISBN network and the International DOI Foundation as their preferred metadata standard for descriptive data associated with the International Standard Book Number and Digital Object Identifier[5].

A study was commissioned recently to examine the potential for ONIX to be used in the library sector[6]. The fact that ONIX is based on the IFLA Functional Requirements for Bibliographic Records model means that there is basic compatibility with library applications. Indeed the trade sector can be commended for not trying to reinvent the wheel and for drawing on the best of what the library world can offer. Librarians who were consulted as part of the study were enthused by the potential for trade-library collaboration, and saw ONIX as having tremendous potential for acquisitions functions and for enriching catalogue records. Indeed some foresaw ONIX as having the potential eventually to replace MARC as the format for carrying cataloguing data. Obviously extension to wider library applications will require more data elements, e.g. preservation data, but the format is expansible.

This raises the question for libraries - which data dictionary does one choose? Are Dublin Core and ONIX competing or complementary? For a time there was concern that they were competing. However, the messages one hears more recently are of convergence, e.g. with the adoption of qualifiers by DC, and of dialogue between the communities. Which one predominates in the library sector will depend on the capabilities of library systems, on the availability of records in the different formats and on the investment that is put in

to developing the systems and schemes. In this situation it is obvious that interoperability will be paramount, and that realisation is gaining ground.

FORMATS

The pre-existing format in widespread use in libraries is MARC. It is well established in the library sector, but not so well in the trade sector. This makes cross sector collaboration more difficult.

The main drawback of MARC is the number of national variants and the apparent competition between UNIMARC and MARC21. The difficulties of exchanging records are pushing countries like the UK and organisations like LIBER to pursue harmonisation of formats. The predominance of the US in research - and to a degree in publishing - means that the target format of choice is becoming MARC21.

The UK is almost certainly going to harmonise its UKMARC format with MARC21, provided the role for non-US organisations in the governance of MARC21 satisfies the UK community[7]. And the future for harmonisation by LIBER members looks promising, though there are changes in the format to accommodate European practice that have to be negotiated. Harmonisation of MARC formats is the topic of another paper by this author presented to the LIBER Annual Conference[8].

XML[9], the major new format that carries out the same function as MARC in a digital environment, grew out of the Standard Generalised Markup Language (SGML) originally developed to support computer typesetting, and the Hypertext Markup Language (HTML) developed for writing Web pages. Combined with Java script XML is an even more powerful web authoring language.

RDF, the Resource Description Format[10], was developed by the World Wide Web Consortium (W3C) as a metadata format designed to be held within XML documents. XML/RDF covers technical and rights data as well as the description of the digital document. Both Dublin Core and ONIX records can be represented in XML/RDF; indeed ONIX was explicitly designed for use in an XML environment.

Two interesting features of RDF are that it tracks the sources of individual data elements - very useful when records are built up by a variety of organisations enhancing data along the line – and that it can provide an envelope for data in other formats. So it would be possible, say, for a library supplier to

embed a MARC record for library use in a publisher's metadata record. The potential implications of that are explored below.

WHAT IS THE FUTURE FOR MARC?

This question took on a special significance for the UK community when it began to consider embarking on harmonising its national format with MARC21. It would be especially unfortunate if libraries were to invest in changing their systems to MARC21 only to find that within a few years MARC is replaced, say, by ONIX or Dublin Core records encoded in the XML format. As leader of the harmonisation consultation exercise I approached three ‚gurus' of the information world for their opinion on the future of MARC. I shall not disclose the identities of the gurus, in case I have oversimplified what they told me, or in case they have since modified their views! Interestingly no single common view emerged.

One leading light, closely involved with developing the MARC format, saw MARC lasting at least 15 years in to the future and probably longer, because of the huge investment by libraries in systems and data. Crosswalks between formats were seen as the way of achieving interoperability, and continuing development of the format would accommodate new applications and data types.

Another guru closely associated with Dublin Core believed that the sheer weight of economics would force libraries to cut cataloguing costs, as they have to cope with increasing volumes of material in both print and digital forms. Because the simple structure of Dublin Core makes it cheap to apply, he predicted that DC would replace MARC for cataloguing both digital and print documents.

The third opinion former, who is not associated with any particular standards camp, postulated that XML format records would be used cross-sectorally for carrying data, but that MARC records for libraries might be embedded within the envelope of the XML records. In that way libraries could continue to capitalise on their existing systems and catalogue databases while reaping the benefits of interoperability and the rich data being generated in the trade sector.

The conclusion drawn as a result of these divergent views and other soundings was that MARC harmonisation is still a valid goal, not only for its immediate benefits for record sharing across international boundaries, but also to

provide a common platform from which libraries could make a coordinated jump to a new standard in the future, saving costs through having common conversion tools.

CATALOGUING RULES

Cataloguing rules are the area where there is arguably the least international standardisation. This could be a barrier that remains when format issues etc are resolved. The Anglo-American Cataloguing Rules (AACR) is the most widely used code, but its influence is confined largely to the English-speaking world. However, the growing interest in AACR from other regions, including LIBER members, has prompted a review by the Committee of Principals for AACR, which is looking at internationalisation of the code's development and governance.

The virtual library environment means that not only must cataloguing rules cope with digital documents, but they must be also be capable of facilitating access to items from a collection that now expands well beyond the physical confines of one institution. The Committee of Principals held an international conference in 1997[11] to discuss how AACR might have to change for the new environment. The recommendations of the conference shaped the development process that has been going on ever since[12].

One of the major conference recommendations was that there needed to be a new logical data model underpinning the standard, to provide a solid foundation upon which to modify and extend the structure. The obvious choice for the data model was the IFLA Functional Requirements for Bibliographic Records (FRBR)[13], and work is going on to fit the Rules to the model. This sounds quite radical, but it is more a question of rearranging the Rules rather than changing them. The aim is to avoid a step change like there was from AACR1 to AACR2, which required substantial conversion of existing catalogues.

The most significant changes are in the definition and cataloguing of serials. The situation has already been complicated enough in the print world with three codes in use: AACR, the IFLA International Standards Bibliographic Description for Serials - ISBD(S) - and UNESCO's International Serials Data System (ISDS). Fortunately, there has been a major breakthrough with the recent agreement to harmonise the codes. In parallel with this harmonisation the definition of what constitutes a serial has been substantially revised, and it

is now much broader to reflect the changing patterns of publication, especially in the digital world.

Another area of change being considered is whether one continues to catalogue the work in hand (currently prescribed by AACR) or the intellectual work, where the physical manifestation acquired by the library is recorded at a secondary level. This has been characterised as the „content versus carrier" debate. The FRBR favours a tiered approach working from the intellectual work at the highest level, as does ONIX, since it has adopted the FRBR model.

So work is well advanced to adapt at least one major cataloguing code to a hybrid print and digital environment. What's more, provided AACR does increase its international appeal by being more welcoming to non-English speaking communities, it may offer a common code in which catalogue data can be more readily exchanged. However, the difficulty of overcoming cultural differences between national codes should not be underestimated.

DRIVERS FOR CHANGE

To begin to answer the questions posed at the beginning of this paper one needs to be aware of the drivers for change.

Within the trade in particular e-commerce is forcing the pace. Amazon and bol.com rely on their catalogues to sell books. They want more and more information in standard form from publishers. In the library world suppliers are offering discounts if libraries use EDI for ordering, chasing and billing functions, as this saves the suppliers money.

Economic pressures are always with us. Cataloguing is often singled out for scrutiny in trying to pare costs. Cataloguing has adapted to these pressures by collaboration, and there are many success stories, e.g. OCLC, the Research Libraries Group and the Consortium of European Research Libraries. However, even more collaboration will be needed to contain costs in the digital environment. Indeed, as libraries try to cope with constrained acquisitions and operating budgets, collaboration becomes a major driver in itself. For instance in the UK university and national library sectors much work is being devoted to developing collaborative acquisition, retention and preservation schemes, and cataloguing systems need to be able to cope with the demands this collaboration places upon them.

Collaboration is also a major force in the trade sector. Book publishers, music publishers and video companies want to ensure their products are readily available from the on-line retailers. So they, too, are collaborating.

For collaboration to work interoperability is paramount. The trade does not want to have to re-key data between applications. Nor do libraries want to convert records between formats as they exchange records.

Lastly, but of greatest importance for a service industry, are the users' and their needs. Users want a hybrid library; they don't want to search separate catalogues for print and digital documents. Neither are they interested in the distinction between what their library holds locally and what is on Web. So they are driving libraries towards providing seamless services, where again interoperability is the key.

ISSUES FOR THE LIBRARY SECTOR

There are a number of issues for the library community that also have to be taken into account when weighing up the answers to the questions about the future.

The first is the systems support for the new formats. We are seeing hybrid systems beginning to emerge from some of the larger vendors. Support for XML will improve as the database management systems that underpin many of the library vendors' packages improve their XML support. For example Oracle, which underpins several popular packages, supports XML and is introducing more features with each new release, e.g. a better nesting capability.

The cost of creating metadata is a major issue. Preliminary results of trials by the British Library show that the effort to create a metadata record for a digital publication can take up to three times longer than cataloguing a printed book. There is all the technical data that has to be captured, which is essential for access to the electronic texts and for preservation. Indeed the ideal time to make preservation decisions is at the time of cataloguing, so that the decisions can be recorded. Then there are the time penalties of having to load digital documents, especially from DVD, CD-ROM etc, in order to discover the technical and descriptive data. All this slows the process.

Record transfer and other forms of collaboration have already been mentioned as a way of minimising and sharing those costs. Another parallel strategy is the creation of productivity tools to gather the information automa-

tically. An example of this approach is the range of productivity tools developed and being refined as part of the CORC (Cooperative Online Resource Catalogue) programme initiated and led by OCLC.

Another issue for libraries are what records are available, at what cost and whether they are of useful quality. A traditional problem at the trade/library interface has been the quality of records produced by publishers, whose staff in many cases do not have appropriate training. This is still an issue to a degree, but, whatever one thinks about the descriptive cataloguing data provided by publishers, there is still value in capturing the rich additional data produced by them. And, of course, the increasing role of commercial bibliographic agencies is helping to provide cleaner descriptive data, too.

The last major issue, especially if common standards prove an elusive goal between different sectors, is the interoperability of standards, in particular the availability of crosswalks. It is encouraging to note that these are being developed, for instance between MARC and ONIX.

THE ANSWERS – OR NOT

Has this analysis of the drivers and issues got us any closer to answering the questions posed at the beginning?

The wide variety of views from a number of people in key roles in the information sector have already been highlighted. Anyone with a reasonable overview of the situation could come up with a valid prediction of the future that might be just as likely to be proved right as the widely varying predictions of the gurus. In order to provoke debate I shall climb out on a limb by giving you my opinion. I do not pretend that it is likely to be any more right than those of my more illustrious colleagues, it is merely one opinion among many. As a safeguard, though, I shall temper my predictions with a degree of caution to salvage some credibility if I am subsequently proved wrong.

What standards will emerge? I do not think we shall see one standard making a clean sweep across all sectors and applications, but a co-existence. Some standards will find niches where they are best suited to the immediate application. Nevertheless we shall see increasing convergence towards a few key standards, and improved interoperability between those standards, but no single overall „winner".

Does MARC have a future? Yes, but finite. The big issue is for how long. Inertia favours a longer life for it. On the other hand libraries have to replace systems from time to time, and at that stage data conversion is often required - so there are opportunities for change. However, libraries need to some degree to move forward together to keep their collaboration schemes functioning. Co-ordinating that step will be difficult and require leadership by the national libraries.

Shall we see print and digital documents being catalogued according to the same - rather than medium specific - standards? Probably. There are those that say it is not essential if we have sophisticated hybrid systems capable of searching multiple databases as if they are one. However, I believe the discontinuities at the boundaries, as have been encountered in existing applications such as Z39.50 searching, will tend to favour common standards covering both print and digital media.

Lastly will we see convergence between libraries and the trade? I think economic pressures and common goals in controlling the e-commerce environment will make that an almost certain outcome, and this is perhaps the most far-reaching change we are likely to see.

Crystal ball gazing is notoriously difficult, and there are probably many different views that have as much if not a greater validity. Even if the reader disagrees with the views expressed, this paper will have served a purpose if it stimulates a debate that helps to get us collectively closer to the answers.

References

1 <http://www.dublincore.org/>.

2 <http://www.editeur.org/onix.html>.

3 <http://www.bic.org.uk/ddinfo.html>.

4 <http://www.indecs.org/project.htm>.

5 <http://www.doi.org/>.

6 <http://www.bic.org.uk/onixlibrep.doc>.

7 All change for UK library cataloguing format – 19 March 2001, <http://www.bl.uk/pr2001/14.htm>.

8 Harmonisation of MARC and descriptive cataloguing standards. Stuart Ede. LIBER Quarterly Vol. 11 (2001), No. 4, pp. 345-353.

9 <http://www.w3c.org/XML/>.

10 <http://www.w3c.org/RDF/>.

11 The principles and future of AACR : proceedings of the International Conference on the Principles and Future Development of AACR : Toronto, Ontario, Canada, October 23/25, 1997 / edited by Jean Wiehs. — Ottawa : Canadian Library Association ; London : Library Association Publishing ; Chicago : American Library Association, 1998. — ISBN 0-88802-287-5 (ALA); 1-85604-303-7 (LA)

12 <http://www.nlc-bnc.ca/jsc/current>.

13 Functional requirements for bibliographic records: final report. UBCIM Publications - New Series Vol 19. IFLA Study Group on the Functional Requirements for Bibliographic Records (September 1997). K G Saur, München 1998. <http://www.ifla.org/VII/s13/frbr/frbr.htm>.